2012

Structural Variation in the Maize Genome

Kai Ying
Iowa State University

Structural variation in the maize genome

by

Kai Ying

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Genetics
(Computational Molecular Biology)

Program of Study Committee:
Patrick S. Schnable, Major Professor
Dan Nettleton
Srinivas Aluru
Thomas Peterson
Xun Gu

Iowa State University

Ames, Iowa

2012

TABLE OF CONTENTS

CHAPTER 1. GENERAL INTRODUCTION

Introduction

Introduction of maize

Maize (*Zea mays* L. *ssp. mays*) is one of the world's most important crops. Maize was domesticated from *Zea mays* L. *ssp. parviglumis* approximately 10,000 years ago in the Balsas River Basin of southwestern Mexico(Hufford et al., 2012). During domestication significant phenotypic transformations (e.g., tiller number, seed architecture, and structures of inflorescences) occurred as compared to the wild ancestor. After domestication, maize has been subjected to a continuous breeding history to gradually improve its yield and other agronomic traits, including modern breeding efforts to produce hybrid maize lines.

Review of Genome Structural Variation (SV)

Structural Variation (SV) is defined as genomic variation affecting a sequence of more than 1 kb in length including insertions/deletions(INDELs), duplications, inversions, translocation and other genome rearrangements (Feuk, Carson, & Scherer, 2006). SV has been detected in many genomes and is often extensive. For example, it is estimated that up to 12% of the human genome is subject to SV (Redon et al., 2006). SV is not evenly distributed in genome; regions with high and low frequencies of SVs exist in both major crops and model plants (Pang et al., 2010).

SV is thought to be an important factor in determining phenotypic variation for a wide range of traits (Stankiewicz and Lupski, 2010). In humans, SV has been reported to be associated with autism, schizophrenia and Crohn's disease (Baker, 2012). In plants, SV has been hypothesized to be a driving force behind phenotypic variation (Chia et al., 2012) . Copy Number Variation (CNV), a type of SV that certain regions of genome have abnormal number of copies, sometimes exhibit strong associations with specific biological functions. Early in 1991, Lupski et al. found the first case that the hereditary neuropathy Charcot-Marie-Tooth disease is associated not with an abnormal version of a particular genetic region but instead with the presence of an extra copy of the normal version (Lupski et al., 1991). Subsequently, numerous cases of CNVs for specific genes or gene families have been associated with different traits. Examples from plants include: boron toxicity tolerance and winter hardiness in barley (Sudmant et al., 2010; Sutton et al., 2007) and dwarfism and flowering time in wheat (Díaz-Castillo, Xia, & Ranz, 2012). In Arabidopsis it has been reported that under stress conditions, after 5 generations of culture, novel CNVs are observed (DeBolt, 2010),which suggests that CNVs can rapidly emerge and be fixed by various selective processes such as breeding if they have positive phenotypic effect.

Overview of Structural Variation in Maize Genome

Maize is a species with extraordinary levels of diversity at the genomic (Buckler, Gaut, & McMullen, 2006), transcriptomic (Flint-Garcia et al., 2005; Stupar & Springer, 2006; Swanson-Wagner et al., 2006) and phenotypic (Buckler et al., 2006; Flint-Garcia et al., 2005) levels. At the genomic level, it exhibits a high level of Single Nucleotide Polymorphisms (SNPs; Vroh Bi et al., 2006), INDEL Polymorphisms (IDPs; Fu et al., 2006), and SV (Buckler et al., 2006; Messing & Dooner, 2006). The frequency of SNPs among maize inbreds is about 10 times higher than that in human. In 2009, the completion of the first draft sequence of maize genome (Schnable et al., 2009) was the starting point for systematic studies of genome variation among maize lines. Initially, these studies were focused on the high throughput discovery of SNPs and their application as markers for Genome-wide Association Studies (Gore et al., 2009). But growing evidence showed that SV can affect phenotypic variation. The majority of SVs within maize are duplications, deletions and insertions among inbreds (Scherer et al., 2007). Nearly 85% of the maize genome is covered by transposable elements (Schnable et al., 2009). The activity of these transposable elements can shuffle the gene space, resulting in CNV (both gains and losses). A special case is the deletion of unique regions in chromosomes. Those regions that are present in one haplotype but entirely missing in other haplotypes are termed as presence-absence variation (PAV; Springer et al., 2009).

Array CGH Based approaches to study Structural Variation

The introduction of array-based technologies has opened the door to large-scale assessments of the complexity of SV. Array-CGH is the comparative hybridization of two labeled samples, usually a test sample and a reference sample, to a set of hybridization targets (probes) on an array. The ratio of hybridization signals obtained at each probe from the two samples is related to copy number. Consequently, the ratio of hybridization signals of the two samples can be used to estimate *relative* copy number differences.

Since the first publication using CGH (Kallioniemi et al., 1992) and the first genome-wide array CGH analysis of genomic DNA (Snijders et al., 2001). Array-based screening has replaced karyotyping as the primary tool for the study of CNVs. It can analyze large number of samples at lower cost, which is essential for sampling the large population sizes necessary for Genome-Wide Association Study (GWAS). There are two types of major commercial DNA arrays-- CGH arrays with 50-100 bp long oligo-nucleotide probes and SNP arrays with <30 bp short probes. Both arrays target the non-repeat regions of sequenced genomes or transcriptomes. Both long-oligo CGH arrays and short-oligo SNP arrays can be used for the study of SV. CGH arrays are hybridized with two labeled

samples (channels) per microarray, while SNP arrays usually are hybridized with only a single sample (channel) per microarray. Both CGH arrays and SNP arrays combine information from multiple probes to call SV. Compared to short-oligo SNP arrays, long-oligo CGH arrays have higher signal-to-noise ratio per probe. Generally, the longer oligo nucleotide probes on CGH arrays are only sensitive to large sequence variations such as INDELs or CNVs, while SNP arrays are designed to detect previously defined SNPs located at exactly in the middle of probe sequences. Commercial and custom CGH & SNP arrays can contain tens of millions of probes and detect variation at Mb to Kb resolution.

Array based genotyping and mapping

One of the goals of functional genomics is to understand the function of each individual gene. This can be achieved via at least two approaches: forward genetics and reverse genetics (Candela and Hake 2008; Jander et al. 2002; Peters et al. 2003). Forward genetics attempts to identify genes and other genomic elements that when mutated result in a particular phenotype. In contrast, reverse genetics attempts to discover the function of a given gene or other genomic element by analyzing the phenotype that results following its mutation.

Forward genetics tries to identify the locus responsible for a phenotype by comparing the DNA of mutant individuals with non-mutant individuals. Mapping the causative locus can be achieved via linkage mapping and or association mapping. In either approach the relationship between phenotype and the causal genomic variation is determined by identifying correlations between genetic markers and phenotypic scores. Genetic markers can be defined as "heritable polymorphisms that can be measured in one or more populations of individuals"(Davey et al., 2011). They are widely used in areas such as population genetics, quantitative genetics, ecological genetics and evolutionary genetics. Ideal genetic markers are evenly distributed across the genome and can be measured easily and inexpensively (Luikart, England, Tallmon, Jordan, & Taberlet, 2003). Traditional genetic markers, such as microsatellites (Jarne & Lagoda, 1996), RFLPs (Botstein, White, Skolnick, & Davis, 1980) and AFLPs (Vos et al., 1995), which usually measured by gel electrophoresis of PCR products, have several clear shortcomings: (1) They include cost and time-consuming cloning and primer design steps. (2) When applied to large populations, genotyping is expensive and labor intensive. With the development of microarray technology, especially commercial high throughput SNP genotyping arrays, the genotyping bottleneck was removed. SNP genotyping arrays offer a distinct advantage in terms of throughput and cost. But at the design stage, the discovery of SNP markers and validation of chip is still tedious and usually requires several iterations. Using CGH arrays, it is not necessary to have prior knowledge of polymorphisms. A whole genome CGH array can be designed using only reference genome sequences. Polymorphic probes will be discovered based on the results of

hybridization experiments. Optionally, these polymorphic probes can be selected to develop a new array for genotyping purposes.

The Origin of CNVs/PAVs

There are a number of potential mechanisms for the *de novo* formation of CNVs/PAVs including non-allelic homologous recombination (NAHR), rearrangements associated DNA repair by non-homologous end-joining (NHEJ), micro-homology mediated break-induced replication (MMBIR), contraction or expansion of variable number tandem repeats (VNTRs) and mobile element insertions (MEI) (for review see Hastings, Lupski, Rosenberg, & Ira, 2009). These can in general be categorized into recombination-based and replication-based mechanisms (Innan & Kondrashov, 2010). For recombination-based mechanisms both homologous recombination (HR) and non-homologous recombination may generate CNVs/PAVs. Non-allelic homologous recombination (NAHR), which occurs due to aberrant pairing of regions of extended homology, is an important source of CNVs/PAVs. Either double-strand break (DSB) repair or break-induced replication (BIR) can induce NAHR. In addition to recombination-mediated mechanisms, replication errors, such as slippage at variable numbers of tandem repeat (VNTR) loci or insertion of transposable elements, also can generate CNVs/PAVs. Both recombination or replication mediated CNV/PAV mechanisms are related to certain kind of recombination or replication "error" and corresponding error repair mechanisms. One consequence of those mechanisms is that CNV/PAV formation appears to occur at higher rates in certain genomic regions (hotspots). In particular, CNVs/PAVs associated with NAHR preferably occurred in regions have local sequence homology (Shao et al., 2008).

A change in copy number requires a change in chromosome structure, joining two formerly separated DNA sequences. Careful analysis of the junction sequences or breakpoints may give important insights about how the structural changes are caused. For example in human genome duplications are more likely to be formed by NAHR, VNTR and retro-transposition.

CNVs are not randomly distributed in the genome; they tend to be clustered together and are found to co-localize with certain low copy repeats (LCRs). LCRs are sequences that occur only one or two times in a haplotype. They can provide the homology needed for recombination. Those CNV/PAV-rich regions can form complex genomic architecture, which consist of CNV genes surrounded by direct and inverted LCRs.

Biological effects of CNVs/PAVs The biological effects of CNVs/PAVs are dependent on the affected sequences and their interactions with the rest of the genome. It is expected that the relative importance of CNVs/PAVs will be higher if they contain regulatory regions and/or genes. For example, in tomatoes an insertion of a 6–8 kb transcription factor dramatically influence tomato fruit (Cong, Barrero, & Tanksley, 2008). CNVs/PAVs also have a considerable impact on plant phenotypes, especially such biological process as pathogen response and heterosis. In plants, genes that function in environmental responses such as stress and disease resistant are usually not essential for viability. They tend to form large gene families consisting of many paralogous genes. The number of members in a family in different individuals/lines may differ (CNVs/PAVs).

Perhaps the most important achievement of modern maize breeding was the discovery of the heterosis. In certain combination of two maize inbred lines, their heterozygous F1-progeny are superior vigor than both of their homozygous parents. Surveys have revealed that two parental lines contain different copies of genes (CNVs), or in some special cases, some genes only exist in one of the parents (PAVs). Expression profile studies show that those genes show CNVs/PAVs in two parents may complement each other in the F1 (Garcia et al. 2008), which causes the F1 to express more genes than either of the two parents. This kind of expression level superior may contribute to hybridize superior vigor (Paschold et al., 2012). The different parents of inbred lines that show heterosis can be further cluster into heterotic groups. Some of those PAVs/CNVs are common within certain heterotic group, which is at least consistent with the hypothesis that complementation of PAVs/CNVs may contribute to heterosis.

Additional copies of a gene may or may not change the overall expression levels of that gene. There are several models to explain the expression patterns of CNV genes. In the neutral model, the extra copies of genes are under no selection or weak purifying selection and seem to be non-adaptive or disadvantageous. But in some cases extra copies of a gene may increase expression levels as compared to a single copy haplotype (dosage effect). While in the pseudo gene model, the extra copies of a gene soon became pseudo genes and are rapidly purged from the population under purifying selection, and they are present now only because they has not yet been purged. In this case the extra copies of genes are usually not expressed. In the neo-function model, additional copies of genes provide redundancy for the original essential copy of the gene, allowing the extra copies of genes to evolve new or modified functions, potentially with novel expression patterns while the original copy retains its original function and expression pattern. The new copy of a gene under recent positive selection usually exhibits a high ratio of non-synonymous vs. synonymous mutations (Inoue & Lupski, 2002).

Research goals

Develop an array-CGH based platform for the detection of SVs (CNVs/PAVs) among maize lines

Microarray-based comparative genomic hybridization (array-CGH) has proved useful in detecting genomic variation within species such as human (Iafrate et al., 2004), chimpanzee (Perry et al., 2008), rat (Guryev et al., 2008), mouse (Lakshmi et al., 2006), dog (Chen, Swartz, Rush, & Alvarez, 2009), and bovine (Fadista, Thomsen, Holm, & Bendixen, 2010). But at the time these studies were initiated high throughput SV detection tools were still lacking for plants. Our goals for this research included: (1) Designing a high-density long-oligo nucleotide array based on the newly available maize genome sequence. (2) Establishing an experimental protocol for DNA sample extraction, labeling and comparative hybridization for maize and teosinte samples. (3) Establishing a computational pipeline and statistical model for array-CGH data normalization, model fitting and reliable CNV/PAV calling. (4) Cross validating array-CGH results with other CNV detection methods such as real-time PCR (Chapter 2).

Generate a systematic description of the structure and distribution of SVs (PAVs & CNVs) among maize inbreds (focusing on the specific case of B73 and Mo17)

There are already numerous reports about SVs (PAVs & CNVs) in human disease and animal breeding. But our knowledge about the genome-wide distribution of SV in plants is still limited. Using the array CGH, we sought to: (1) identify a reliable set of CNVs/PAVs in maize two inbred lines: B73 & Mo17; (2) summarize the length, genomic distribution and frequency of CNVs between the inbred lines B73/Mo17; (3) identify some significant characteristics such as hot spots of SVs and highly conserved regions; (4) survey the population structure of selected SVs. (Chapter 2)

Study the expression pattern of SVs (PAVs & CNVs) related genes among different maize inbred lines (in special case between maize inbred line B73 & Mo17)

There are different models for the fates of SV related genes such as the neutral model, the pseudo-gene model, and the neo-function model. The different models predict different expression patterns for the related genes. Survey of those expression patterns may help us to understand their underlying evolutionary patterns and potential biological functions (Chapter 2).

Apply SVs as new markers for genotyping

Building a cost efficient, time and labor saving genotyping platform would be valuable. Based on array-CGH results of maize inbred lines B73, Mo17 and two recombination inbred lines (RILs) of

the Inter-mated B73 x Mo17 (IBM) population, we tried to: (1) select a subset of probes that show high signal difference between the two parental lines (B73 and Mo17) to build a genotyping array; (2) confirm that hybridizing signals of RILs can be used to determine the original source of its genomic segments; (3) establish a reliable, robust and simple computational data process pipeline for genotyping using CGH data(Chapter 3).

Explore potential mechanisms for the origin of CNVs/PAVs

SV (CNVs & PAVs) has been reported to arise via various mechanisms. Most examined cases are, however, from human or other animals that have relative low levels of intra-species diversity. After the creation of SVs, fixation of SVs in species/sub-species is a continue going process. For crops such as maize, which has a long history of domestication and breeding the mechanisms by which SV arises and is fixed is still not clear. Analysis of the structure of CNVs/PAVs and flanking regions may provide important insight into the mechanisms by which they arise. In different maize lines, there exist numerous single copy non-allelic homologs, the segregation of these loci in the descendants of an F1 will yield losses and gains of pairs of non-allelic homologs. Analysis of Array-CGH data provided support for this hypothesis. Additional Next Generation Sequence and other experimental methods were used to test this hypothesis(Chapter 4).

Dissertation organization

This thesis is a collection of the research on SV conducted during my PhD studies. Prior to the release of the maize genome sequence, our lab had already discovered nearly identical paralogs (NIPs), duplicated gene segments that exhibit very high levels of sequence similarity (Emrich et al., 2007). With the availability of the maize genome in 2009, I had the opportunity to systematically identify hundreds of NIPs; the results of these analyses were published as part of maize genome sequencing paper(Appendix A). To identify CNV among maize lines, new technologies such as Array-CGH were required. Chapter 2 is a report of an analysis of CNVs/PAVs in two maize inbred lines (B73 & Mo17) using array-CGH technology. This work is the outcome of a collaboration of our lab with Drs. Nathan Springer of the University of Minnesota and Jeffrey Jeddeloh of Roche Nimblegen. The whole project was under the direction of Dr. Schnable. Most of the research was conducted in 2007-2008 and the paper was published in PLoS-Genetics as a “companion paper” with the major maize genome paper (Schnable et al., 2009). This is one of the first papers that apply array-CGH technology to a

major crop. I conducted most of the bioinformatics analyses, including array-CGH normalization, segmentation and PAV/CNV calling. In this paper for the first time we revealed that maize has high level of CNVs/PAVs, which is an important aspect of its genome structure and may contribute to the phenotypic diversity of this important crop. The high level of SV among maize lines and its potential relation with phenotype diversity interested me and induced me to select SV of maize genome as the major research topic for my PhD research.

To extend the application of array-CGH technology in comparative genomics and explore array-CGH potential as a genotyping tool, CGH experiments were performed on two lines of the maize Intermated B73 x Mo17 (IBM) mapping population. The results show that the majority of probes that have significant differences in hybridization signals between two parental lines (B73 & Mo17) also have similar differences in RILs. Hence, array-CGH signals can be used as genetic markers for genotyping purpose. This work is presented in chapter 3 as a joint paper with Drs. Yan Fu and Nathan Springer (Fu et al., 2010). I was responsible for most of the array-CGH bioinformatics analyses.

Detailed analyses of the hybridization results from the IBM RILs revealed that some of the probes exhibit hybridization signals that differ from those of both parents. One of my lab mates, Dr. Sanzhen Liu, hypothesized that this was caused by the segregation of non-allelic homologs. Various experiments were designed and conducted that in the end provided substantial support for this hypothesis and ruled out other potential explanations. This led to the discovery of a "novel" mechanism by which CNVs/PAVs can originate. This report is provided in chapter 4 of thesis and was published in plant journal(Liu et al., 2012). I was responsible for the bioinformatics analyses of the array-CGH data and the pair end sequence analyses.

References

Baker, M. (2012). Structural variation: the genome's hidden architecture. *Nature methods*, *9*(2), 133–7. doi:10.1038/nmeth.1858

Botstein, D., White, R. L., Skolnick, M., & Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American journal of human genetics*, *32*(3), 314–31. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1686077&tool=pmcentrez&rendertype=abstract

Buckler, E. S., Gaut, B. S., & McMullen, M. D. (2006). Molecular and functional diversity of maize. *Current opinion in plant biology*, *9*(2), 172–6. doi:10.1016/j.pbi.2006.01.013

Chen, W.-K., Swartz, J. D., Rush, L. J., & Alvarez, C. E. (2009). Mapping DNA structural variation in dogs. *Genome research*, *19*(3), 500–9. doi:10.1101/gr.083741.108

Chia, J.-M., Song, C., Bradbury, P. J., Costich, D., de Leon, N., Doebley, J., Elshire, R. J., et al. (2012). Maize HapMap2 identifies extant variation from a genome in flux. *Nature genetics*, *44*(7), 803–7. doi:10.1038/ng.2313

Cong, B., Barrero, L. S., & Tanksley, S. D. (2008). Regulatory change in YABBY-like transcription factor led to evolution of extreme fruit size during tomato domestication. *Nature genetics*, *40*(6), 800–4. doi:10.1038/ng.144

Davey, J. W., Hohenlohe, P. a, Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature reviews. Genetics*, *12*(7), 499–510. doi:10.1038/nrg3012

DeBolt, S. (2010). Copy number variation shapes genome diversity in Arabidopsis over immediate family generational scales. *Genome biology and evolution*, *2*, 441–53. doi:10.1093/gbe/evq033

Díaz-Castillo, C., Xia, X.-Q., & Ranz, J. M. (2012). Evaluation of the role of functional constraints on the integrity of an ultraconserved region in the genus Drosophila. *PLoS genetics*, *8*(2), e1002475. doi:10.1371/journal.pgen.1002475

Emrich, S. J., Li, L., Wen, T.-J., Yandeau-Nelson, M. D., Fu, Y., Guo, L., Chou, H.-H., et al. (2007). Nearly identical paralogs: implications for maize (Zea mays L.) genome evolution. *Genetics*, *175*(1), 429–39. doi:10.1534/genetics.106.064006

Fadista, J., Thomsen, B., Holm, L.-E., & Bendixen, C. (2010). Copy number variation in the bovine genome. *BMC genomics*, *11*, 284. doi:10.1186/1471-2164-11-284

Feuk, L., Carson, A. R., & Scherer, S. W. (2006). Structural variation in the human genome. *Nature reviews. Genetics*, *7*(2), 85–97. doi:10.1038/nrg1767

Flint-Garcia, S. a, Thuillet, A.-C., Yu, J., Pressoir, G., Romero, S. M., Mitchell, S. E., Doebley, J., et al. (2005). Maize association population: a high-resolution platform for quantitative trait locus dissection. *The Plant journal : for cell and molecular biology*, *44*(6), 1054–64. doi:10.1111/j.1365-313X.2005.02591.x

Fu, Y., Springer, N. M., Ying, K., Yeh, C.-T., Iniguez, a L., Richmond, T., Wu, W., et al. (2010). High-resolution genotyping via whole genome hybridizations to microarrays containing long oligonucleotide probes. *PloS one*, *5*(12), e14178. doi:10.1371/journal.pone.0014178

Fu, Y., Wen, T.-J., Ronin, Y. I., Chen, H. D., Guo, L., Mester, D. I., Yang, Y., et al. (2006). Genetic dissection of intermated recombinant inbred lines using a new genetic map of maize. *Genetics*, *174*(3), 1671–83. doi:10.1534/genetics.106.060376

Gore, M. a, Chia, J.-M., Elshire, R. J., Sun, Q., Ersoz, E. S., Hurwitz, B. L., Peiffer, J. a, et al. (2009). A first-generation haplotype map of maize. *Science (New York, N.Y.)*, *326*(5956), 1115–7. doi:10.1126/science.1177837

Guryev, V., Saar, K., Adamovic, T., Verheul, M., van Heesch, S. a a C., Cook, S., Pravenec, M., et al. (2008). Distribution and functional impact of DNA copy number variation in the rat. *Nature genetics*, *40*(5), 538–45. doi:10.1038/ng.141

Hastings, P. J., Lupski, J. R., Rosenberg, S. M., & Ira, G. (2009). Mechanisms of change in gene copy number. *Nature reviews. Genetics*, *10*(8), 551–64. doi:10.1038/nrg2593

Hufford, M. B., Xu, X., van Heerwaarden, J., Pyhäjärvi, T., Chia, J.-M., Cartwright, R. a, Elshire, R. J., et al. (2012). Comparative population genomics of maize domestication and improvement. *Nature genetics*, *44*(7), 808–11. doi:10.1038/ng.2309

Iafrate, a J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., Scherer, S. W., et al. (2004). Detection of large-scale variation in the human genome. *Nature genetics*, *36*(9), 949–51. doi:10.1038/ng1416

Innan, H., & Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. *Nature reviews. Genetics*, *11*(2), 97–108. doi:10.1038/nrg2689

Inoue, K., & Lupski, J. R. (2002). Molecular mechanisms for genomic disorders. *Annual review of genomics and human genetics*, *3*(121), 199–242. doi:10.1146/annurev.genom.3.032802.120023

Jarne, P., & Lagoda, P. J. (1996). Microsatellites, from molecules to populations and back. *Trends in ecology & evolution*, *11*(10), 424–9. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21237902

Kallioniemi, A., Kallioniemi, O. P., Sudar, D., Rutovitz, D., Gray, J. W., Waldman, F., & Pinkel, D. (1992). Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science (New York, N.Y.)*, *258*(5083), 818–21. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1359641

Lakshmi, B., Hall, I. M., Egan, C., Alexander, J., Leotta, A., Healy, J., Zender, L., et al. (2006). Mouse genomic representational oligonucleotide microarray analysis: detection of copy number variations in normal and tumor specimens. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(30), 11234–9. doi:10.1073/pnas.0602984103

Liu, S., Ying, K., Yeh, C., Yang, J., Swanson-wagner, R., Jeddeloh, J. A., Schnable, P. S., et al. (2012). Changes in Genome Content Generated via Segregation of Non-allelic Homologs. *Plant J*. doi:10.1111/j.1365

Luikart, G., England, P. R., Tallmon, D., Jordan, S., & Taberlet, P. (2003). The power and promise of population genomics: from genotyping to genome typing. *Nature reviews. Genetics*, *4*(12), 981–94. doi:10.1038/nrg1226

Lupski, J. R., de Oca-Luna, R. M., Slaugenhaupt, S., Pentao, L., Guzzetta, V., Trask, B. J., Saucedo-Cardenas, O., et al. (1991). DNA duplication associated with Charcot-Marie-Tooth disease type 1A. *Cell*, *66*(2), 219–32. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1677316

Messing, J., & Dooner, H. K. (2006). Organization and variability of the maize genome. *Current opinion in plant biology*, *9*(2), 157–63. doi:10.1016/j.pbi.2006.01.009

Pang, A. W., MacDonald, J. R., Pinto, D., Wei, J., Rafiq, M. a, Conrad, D. F., Park, H., et al. (2010). Towards a comprehensive structural variation map of an individual human genome. *Genome biology*, *11*(5), R52. doi:10.1186/gb-2010-11-5-r52

Paschold, A., Jia, Y., Marcon, C., Lund, S., Larson, N. B., Yeh, C., Ossowski, S., et al. (2012). Complementation contributes to transcriptome complexity in maize (Zea mays L .) hybrids relative to their inbred parents. *Genome research*, 1–26.

Perry, G. H., Yang, F., Marques-Bonet, T., Murphy, C., Fitzgerald, T., Lee, A. S., Hyland, C., et al. (2008). Copy number variation and evolution in humans and chimpanzees. *Genome research*, *18*(11), 1698–710. doi:10.1101/gr.082016.108

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., et al. (2006). Global variation in copy number in the human genome. *Nature*, *444*(7118), 444–54. doi:10.1038/nature05329

Scherer, S. W., Lee, C., Birney, E., Altshuler, D. M., Eichler, E. E., Carter, N. P., Hurles, M. E., et al. (2007). Challenges and standards in integrating surveys of structural variation. *Nature genetics*, *39*(7 Suppl), S7–15. doi:10.1038/ng2093

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., Liang, C., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science (New York, N.Y.)*, *326*(5956), 1112–5. doi:10.1126/science.1178534

Shao, L., Shaw, C. a, Lu, X.-Y., Sahoo, T., Bacino, C. a, Lalani, S. R., Stankiewicz, P., et al. (2008). Identification of chromosome abnormalities in subtelomeric regions by microarray analysis: a study of 5,380 cases. *American journal of medical genetics. Part A*, *146A*(17), 2242–51. doi:10.1002/ajmg.a.32399

Snijders, a M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., et al. (2001). Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature genetics*, *29*(3), 263–4. doi:10.1038/ng754

Springer, N. M., Ying, K., Fu, Y., Ji, T., Yeh, C.-T., Jia, Y., Wu, W., et al. (2009). Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS genetics*, *5*(11), e1000734. doi:10.1371/journal.pgen.1000734

Stupar, R. M., & Springer, N. M. (2006). Cis-transcriptional variation in maize inbred lines B73 and Mo17 leads to additive expression patterns in the F1 hybrid. *Genetics*, *173*(4), 2199–210. doi:10.1534/genetics.106.060699

Sudmant, P. H., Kitzman, J. O., Antonacci, F., Alkan, C., Malig, M., Tsalenko, A., Sampas, N., et al. (2010). Diversity of human copy number variation and multicopy genes. *Science (New York, N.Y.)*, *330*(6004), 641–6. doi:10.1126/science.1197005

Sutton, T., Baumann, U., Hayes, J., Collins, N. C., Shi, B.-J., Schnurbusch, T., Hay, A., et al. (2007). Boron-toxicity tolerance in barley arising from efflux transporter amplification. *Science (New York, N.Y.)*, *318*(5855), 1446–9. doi:10.1126/science.1146853

Swanson-Wagner, R. a, Jia, Y., DeCook, R., Borsuk, L. a, Nettleton, D., & Schnable, P. S. (2006). All possible modes of gene action are observed in a global comparison of gene expression in a maize F1 hybrid and its inbred parents. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(18), 6805–10. doi:10.1073/pnas.0510430103

Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., Frijters, a, et al. (1995). AFLP: a new technique for DNA fingerprinting. *Nucleic acids research*, *23*(21), 4407–14. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10206688

Vroh Bi, I., McMullen, M. D., Sanchez-Villeda, H., Schroeder, S., Gardiner, J., Polacco, M., Soderlund, C., et al. (2006). Single Nucleotide Polymorphisms and Insertion–Deletions for Genetic Markers and Anchoring the Maize Fingerprint Contig Physical Map. *Crop Science*, *46*(1), 12. doi:10.2135/cropsci2004.0706

CHAPTER 2. ARRAY COMPARATIVE GENOMIC HYBRIDIZATION(ARRAY-CGH) REVEAL HIGH LEVELS OF COPY NUMBER VARIATION(CNV) & PRESENT ABSENT VARIATION(PAV) BETWEEN MAIZE INBRED LINES B73 AND MO17.

Modified from a paper published in *PLoS Genetics* 2009, 5(11): e1000734

Nathan M. Springer[1*], Kai Ying[2,3*], Yan Fu[4,5], Tieming Ji[6], Cheng-Ting Yeh[4,7], Yi Jia[8], Wei Wu[4,7], Todd Richmond[9], Jacob Kitzman[9], Heidi Rosenbaum[9], A. Leonardo Iniguez[9], W. Brad Barbazuk[10], Jeffrey A. Jeddeloh[9], Dan Nettleton[6], Patrick S. Schnable[2,3,4,5,7,8**]

[1]Department of Plant Biology, University of Minnesota, Saint Paul MN 55108
[2]Interdepartmental Genetics Graduate Program, [3]Department of Genetics, Development and Cell Biology, [4] Department of Agronomy, [6] Department of Statistics, [7] Center for Plant Genomics, [5] Center for Carbon Capturing Crops, [8] Interdepartment Plant Biology, Iowa State University, Ames, Iowa 50011, [9] Roche NimbleGen, Madison, WI 53719, [10]University of Florida, Gainesville, FL 32610

*These authors contributed equally
** Corresponding author: Roy J. Carver Co-Laboratory Iowa State University Ames, IA 50011-3650.
E-mail: schnable@iastate.edu

Abstract

Following the domestication of maize over the past ~10,000 years, breeders have exploited the extensive genetic diversity of this species to mold its phenotype to meet human needs. The extent of structural variation, including copy number variation (CNV) and presence/absence variation (PAV), which are thought to contribute to the extraordinary phenotypic diversity and plasticity of this important crop, have not been elucidated. Whole-genome array-based comparative genomic hybridization (CGH) revealed a level of structural diversity between the inbred lines B73 and Mo17 that is unprecedented among higher eukaryotes. A detailed analysis of altered segments of DNA conservatively estimate that there are several hundred CNV sequences among the two genotypes, as well as several thousand PAV sequences that are present in B73 but not Mo17. Haplotype-specific PAVs contain hundreds of single-copy, expressed genes that may contribute to heterosis and the extraordinary phenotypic diversity of this important crop.

Author summary

There is a growing appreciation for the role of genome structural variation in creating

phenotypic variation within a species. Comparative genomic hybridization was used to compare the genome structures of two maize inbred lines, B73 and Mo17. The data reinforce the view that maize is a highly polymorphic species but also show that there are often large genomic regions that have little or no variation. We identify several hundred sequences that, while present in both B73 and Mo17, have copy number differences in the two genomes. In addition, there are several thousand sequences, including at least 180 sequences annotated as single-copy genes, that are present in one genome but entirely missing in the other genome. This genome content variation leads to differences in transcript content between inbred lines and likely contributes to phenotypic diversity and heterosis in maize.

Introduction

Although many analyses of genetic variation have focused on single nucleotide polymorphisms (SNPs), there is a growing appreciation for the roles of structural variation as a cause for phenotypic variation [1-7]. Indeed, structural variation can have major phenotypic consequences [6]. The term copy number variation has been used to describe duplications, deletions and insertions among individuals of a species [5]. Herein the term copy number variation (CNV) is reserved to describe sequences that are present in both genomes being compared, albeit in different copy number. The term presence-absence variation (PAV) is used to describe sequences that are present in one genome but entirely missing in the other genome.

Maize is phenotypically diverse [8-9] and this phenotypic diversity is reflected by substantial variation in phenotypic and transcript levels among maize lines [8,10-11]. In addition, the maize genome exhibits extraordinarily high levels of genetic diversity as assayed at the level of SNPs, InDel Polymorphisms (IDPs), and structural variation [9,12]. The frequency of SNPs among maize inbreds is higher than the frequency of SNPs between humans and chimpanzees [9]. The inbred lines B73 and Mo17 are important models for structural and functional genomics of maize. On average, B73 and Mo17 contain an IDP every ~300 bp and SNPs every ~80 bp [13-14] and within transcripts SNPs are found between the inbred lines B73 and Mo17 on average every 300 bp [15]. These levels of diversity are not limited to comparisons between B73 and Mo17. When comparing any two randomly chosen maize inbred lines, there is, on average, one polymorphism every 100 bp [16-17]. Collectively, these studies indicate that maize has relatively high levels of SNPs and IDPs as compared to many other species [9].

There is also cytogenetic evidence for structural variation in the genomes of maize inbreds.

Structural genomic variation involves alterations in DNA sequence beyond SNPs or small IDPs, and includes large-scale differences in chromosomal structure, altered locations of genes or repetitive elements, copy number variation (CNV) and presence/absence differences among haplotypes. Large-scale differences in chromosomal structure between maize inbred lines were first identified through cytogenetic studies. Barbara McClintock and others analyzed heterochromatic knob (highly condensed, tandem repeat regions) content and size to characterize genome variation [18-20]. Recent studies have documented differences in the content of several classes of repetitive DNA between maize inbreds at the chromosomal level [21]. Flow cytometry studies have also documented significant variation in overall genome sizes among inbred lines [22].

Sequence-based methodologies have documented structural diversity at a higher resolution (reviewed by [9,12]). Sequencing of BACs containing the bz1 gene from eight different inbred lines revealed two significant findings [23-24]. First, there is variation for the presence of several genic fragments such that these "genes" are found at this locus in some inbreds but not in others [23]. These "genes" were subsequently found to be gene fragments that had been mobilized by Helitron transposons [25-26]. These are not PAVs because although a genome may lack a copy in the vicinity of the bz1 locus, such a genome typically contained one or more copies of these genes (or gene fragments) elsewhere. Second, comparison of multiple haplotypes revealed major differences in the amount and types of repetitive elements between genes. The same gene can be flanked by very different repetitive elements in different inbred lines [23]. At the same time, similar kinds of repeat diversity between haplotypes were reported in the a1-sh2 interval [27]. Both of these findings have been supported by analyses of other genomic regions in B73 and Mo17 [28]. A study of the presence and location for many genic fragments in B73 and Mo17 BAC libraries suggested that many sequences can vary in location or even presence between B73 and Mo17 [28]. There is also evidence for variation in the presence of nearly identical paralogs (NIPS) in different maize inbred lines [29].

Understanding the intraspecific variation of maize has important implications for crop improvement and plant breeding. Long-term selection experiments have demonstrated a surprising wealth of potential; even when starting with relatively little genetic diversity it has been possible to continue to make phenotypic gains for traits such as oil content for over a century [30]. In addition, the combination of variation from different maize inbred lines in hybrids results in heterosis [31]. The availability of genomic resources for maize, particularly the B73 maize genome sequence [32] has provided an opportunity to conduct genome-wide analyses of structural variation. We have used high-density oligonucleotide microarrays to identify patterns of structural variation across the maize genome. We find evidence for a high rate of CNVs. In addition, we identify several thousand DNA segments, often including genic sequences, that are present in the B73 genome but absent from the Mo17 genome (i.e.,

PAVs). By assessing genome-wide structural variation in maize we have gained a better understanding of the nature of variation among different maize inbred lines.

Results

Development and annotation of a CGH microarray for maize

Genomic variation within a species can be assessed using comparative genomic hybridization (CGH). A high-density (2.1 million feature) oligonucleotide microarray was designed using the sequences of B73 BACs. Probes range in size from 45-85 bp and were selected using slightly relaxed criteria (due to the overlap of adjacent BAC sequences and lack of assembly at the time of design) relative to those traditionally used for CGH probe design. The 2.12M probes were aligned to the B73 RefGen_v1 [32] released by the maize genome sequencing project (MGSP). It was possible to identify perfect matches (100% ID and 100% coverage) for 93% (1.98M/2.10M) of the probes. Approximately ~1.78 million of the probes had a single perfect match and were therefore deemed to be single copy, ~120k probes had two perfect matches and ~34k probes had three perfect matches (Figure S1).

All of these perfectly matched probes were classified based on their repetitiveness and locations relative to predicted genes (see Materials and Methods for details and Table 1 for numbers). Approximately 30% of the probes exhibited evidence of containing repetitive sequences (Methods; Table 1). Probes were also mapped relative to genes and other types of annotation produced by the MGSP. The distributions of probes relative to these types of annotation were assessed by visualizing the locations of probes that aligned to several genomic regions for which high-quality assembled sequence and manual annotation were available for both B73 and Mo17 (Figure 1 and S2; [23,28]. There are 1,604 probes within the ~1 Mb of B73 sequence from these four regions (selected portions are shown in Figure 1 and S2. Probe density is generally high for those regions in which the B73 and Mo17 haplotypes align well. These regions tend to be genic or low copy and have 3-4 probes per kb. Consistent with the NimbleGen probe design strategy, fewer probes are located in regions consisting of a high percentage of repetitive element sequences.

Analyses of these regions were used to evaluate the quality of our genome-wide probe annotation. Several tracks in Figure 1A provide information about the repetitive and genic classifications of probes. Our genome-wide annotations generally agree with the detailed annotation information

available for these four regions. The probes that were classified as repetitive in the genome-wide analyses were often found within sequences that were annotated as repetitive or retrotransposon based in the four regions that had been subjected to manual annotation.

All probes were designed based on the B73 haplotype. To determine whether probe sequences were conserved in Mo17, probe sequences were aligned to a collection of 42,206,644 Mo17 whole-genome shotgun (WGS) reads generated by the DOE's Joint Genome Institute (JGI) and provided to us prior to publication by the Rohksar group. Based on these alignments each probe was classified as being a perfect match (100% identity and coverage), highly conserved (>97% identity and coverage), conserved (>90% identity and coverage), poorly conserved (>75% identity and >70% coverage) or as having no significant match in the JGI Mo17 data set. Over 80% of the probes were at least 90% identical to Mo17 sequences with over 90% of probe sequence coverage (Table 1).

The analysis of the four regions that have complete coverage of the Mo17 haplotype permitted us to compare the results of our genome-wide classifications with actual alignments of complete B73 and Mo17 sequences. Because the JGI collection of Mo17 WGS reads provides approximately 4X coverage of the genome we expect some probes to be mis-classified as poorly conserved or as having no match in Mo17 simply due to incomplete sampling of the Mo17 genome. Overall, there was strong agreement between our classification of probes based on alignments to the Mo17 WGS reads and the genomic alignments shown in Figure 1. As expected, there were few cases of probes within highly conserved regions that had erroneously been classified as poorly conserved or no match. Even so, most probes within regions of the B73 haplotype for which there was no significant similarity in allelic regions of the Mo17 haplotype did not match the WGS Mo17 sequences. Some of the probes that matched regions of the B73 haplotype for which there was no significant similarity in allelic regions of the Mo17 haplotype (i.e., positions 257,000-259,000 in Figure 1A) did have similarity to WGS Mo17 sequences. This suggests that regions of the B73 haplotype that can not be aligned to allelic positions of the Mo17 haplotype are of two types. In some cases the non-aligning sequences are B73-specific (PAVs), while in other cases Mo17 contains these sequences but at non-allelic positions similar to those reported by Fu and Dooner [23].

Structural variation detected by CGH

B73 and Mo17 genomic DNA samples were hybridized to the microarray using dye swaps as well as technical replication (Methods). Analysis of the CGH data reveals a bias towards stronger hybridization signals from B73 genomic DNA than from Mo17 genomic DNA (Figures S3 and 4). This bias is likely due to the fact that the array design was based upon the B73 genomic sequence and that polymorphisms between B73 probes and the labeled Mo17 genomic DNA may reduce signal strength. This imbalance in signals between the genotypes violates an assumption required to perform typical

global normalization (see Supplemental text for further details). Consequently, we implemented a normalization procedure that utilized a subset of probes for array normalization. This strategy employed the raw signals from those 840,289 probes whose sequences are absolutely conserved between B73 and Mo17 (based on our analysis of the Mo17 WGS data) to normalize the remaining data (~60% of the probes). A linear model was used to estimate the signal from each genotype and to determine q-values to control false discovery rates.

To understand the biological causes of differences in hybridization signals between B73 and Mo17 we initially focused on the four regions shown in Figure 1 and S2 for which high-quality B73 and Mo17 sequence were available. We found significant ($q<0.0001$) differences in hybridization signal in B73 relative to Mo17 for 234 of the 1,604 probes within these regions (Table 2). As expected it was much more common to observe higher hybridization signal in B73 (210 probes) than the reverse (24 probes).

There are at least three biological reasons why a probe exhibits significant differences in signal after being hybridized to genomic DNA from two inbred lines. First, the probe sequence may have polymorphisms in the two genotypes (SNPs and IDPs). Second, the copy number of the probe in the genomic DNA might be different in the two genotypes being compared (CNV). Third, the probe sequence may be present in the genomic DNA of the reference genotype but not the other (PAV). It is important to remember that while all three reasons could explain why a probe would have a higher signal in B73 than in Mo17, only the second reason is likely to cause probes to have higher signals in Mo17 than in B73 because all probes were designed based on the B73 sequence.

The impact of sequence polymorphisms on hybridization can be observed by comparing the average $\log_2$(Mo17/B73) in probes with different levels of polymorphism between B73 and Mo17 (Table 2). For probes with no polymorphisms the average $\log_2$(Mo17/B73) is zero. As the number of polymorphisms between B73 and Mo17 increases, the $\log_2$(Mo17/B73) value decreases and the percentage of probes that exhibit statistically significant differences in signal strength ($q<0.0001$) increases. Most of the probes with significant variation (68%) have 5 or more SNPs (note that often these probes cannot be aligned to the Mo17 WGS reads at all and many have multiple IDPs or may even be absent altogether from Mo17). Overall, this finding indicates that the majority of the significant differences in hybridization signals are due to the presence of multiple polymorphisms within the ~70 bp probe sequence or due to sequences that encompass or overlap the probe sequence that are present in B73 but absent from the Mo17 genome.

Further support for the concept that many of the probes that exhibit significant differences in hybridization signals are reporting structural variation was provided by visualization of the distribution

of log2(Mo17/B73) signals relative to the four B73/Mo17 haplotype alignments (Figure 1 and S2). For example, each of the four probes in Figure S2A that have significantly lower signals in Mo17 than in B73 are in regions in which the two haplotypes differ substantially. Similarly, in Figure S2B the six probes with significantly lower signal in Mo17 than in B73 all fall near regions of structural variation. Many of the probes with significant signal differences between B73 and Mo17 occurred in the regions surrounding non-shared repetitive elements. It was surprising that some of the probes with 5 or more SNPs (in alignments between these two regions only) did not exhibit significant differences in hybridization signals. However, we noted that although some of these probes (several examples shown in S2) do not have a similar sequence at an allelic position in Mo17, they do have one present elsewhere in the Mo17 genome based on alignments to the Mo17 WGS sequences.

Genome-wide analysis of probes with variable signal in B73 and Mo17

After analyzing in detail probes that aligned to the regions presented in Figure 1 and S2, we assessed the characteristics of all probes that exhibit significant variation in B73 relative to Mo17. At a cut-off of $q<0.0001$ there are 325,813 probes with significant differences in hybridization signals between B73 and Mo17 (15% of all probes, Table 1). The majority (90%) of these probes exhibit higher signals from B73 than from Mo17 (B>M; Table 1), as can be readily observed in volcano and MA plots (Figure S5). In general, and as expected, repeat probes tend to have higher signals than non-repetitive probes (as seen in MA plots in Figure S6). Both B73>Mo17 and Mo17>B73 probes are enriched for non-repetitive probes and consequently depleted for repetitive probes (Table 1 and Figure S7 and S8A). This is not surprising because the signals associated with repetitive probes reflect cross-hybridization from multiple genomic sites and therefore a change at a single site will have less impact on signal strength.

There are striking differences in the B73>Mo17 and Mo17>B73 probes when comparing annotation based on alignments of probe sequences to the Mo17 WGS sequences (Table 1, Figure S8B, S9 and S10). Consistent with our analysis of probes from Figure 1, genome-wide B73>Mo17 probes are enriched for sequences with no match or poor conservation in the Mo17 WGS sequence and are correspondingly depleted for highly conserved or identical probes. Those B73>Mo17 probes that do have an identical or highly conserved sequence among the Mo17 WGS sequences are likely to be examples of CNV and can be used to estimate the rate of CNV. The Mo17>B73 probes follow the opposite pattern with enrichment for probes that have a highly conserved or identical sequence in both B73 and Mo17. This indicates that many of the probes with no match in the Mo17 WGS sequence reflect actual sequence differences, not simply a lack of coverage in the Mo17 WGS sequence.

Probes were compared to the full “working set” of genes predicted by the MGSP (www.maizesequence.org). This “permissive” gene set (n=129,891) includes low-copy transposons as

well as pseudogenes. The B73>Mo17 probes exhibit a distribution genic and intergenic matches that is very similar to all probes. Interestingly, the Mo17>B73 probes are slightly depleted for intergenic probes and show an enrichment for probes near or within genes (Table 1; Figure S8C). A very similar distribution is observed using the filtered set of high-quality gene annotations from the MGSP (data not shown).

Distribution of structural variation throughout the maize genome

The probes were aligned to the B73 RefGen_v1 to visualize the patterns of structural variation along the B73 and Mo17 chromosomes (Figure 2). It should be noted that while the B73 reference genome generally place segments of DNA in the proper order at the level of a single BAC, the local orientation and order of sequence contigs within a BAC has not always been determined. Therefore, our genomic localization of the probes is likely only accurate within the average size of a BAC (~170 kb). The log2(Mo17/B73) signals for each probe were plotted relative to the genomic localization of the probes. As noted above, the majority of probes with significant B73>Mo17 hybridization detect structural variation. The genomic view provided in Figure 2 reveals that structural variation between these two inbreds is not evenly distributed throughout the maize genome. The large number of data points plotted on this graph (~>2 million) can make it difficult to visualize the relative rates of variation across the genome. Therefore, we implemented a sliding window analysis to observe the frequency of probes with significant B73>Mo17 variation in regions that are the approximately the size of 10 average BACs (Figure 3).

There are a number of highly conserved genomic regions that have very little or no structural variation between B73 and Mo17 (Figure 3). For example, there is an ~19 Mb region on chromosome 8 (Figure S11A; positions 140,904,890 – 158,897,190) and a 17 Mb region on chromosome 1 (positions 121,420,890 – 138,984,608) with no evidence for structural variation. The sliding window analysis identified 104 regions that exhibit little to no structural variation (fewer than 4% of the probes exhibit significant variation). Seven of these low diversity regions (on chromosomes 1, 3, 4, 5, 8 and 9) are over 10 Mb.

We performed further characterization of the large regions on chromosomes 1 and 8 with low rates of structural variation. The majority of probes within these regions (83%) are 100% conserved in B73 and Mo17 suggesting that these are low diversity regions. None of 388 primer pairs designed to amplify sequences within these regions detected sequence variation between B73 and Mo17 that could be detected via agarose gel electrophoresis. In comparison, 13% of all primer sets designed for random genomics sites detect variation. We then used Temperature Gradient Capillary Electrophoresis (TGCE) to test whether B73 and Mo17 amplification products from 156 of the 388 primer pairs from the conserved

regions contain SNPs or small IDPs. TGCE is sensitive enough to detect a single SNP in amplicons of over 800 bp and 1 bp IDPs in amplicons of ~500 bp [33], which is the typical size of these amplification products. Of these 123/156 (79%) exhibited no evidence of even a single SNP or IDP between B73 and Mo17, indicating the high level of sequence conservation within these two intervals. In contrast, only 39% of randomly selected sites are not polymorphic using TCGE assays.

There is a tendency for these large low diversity regions to be located near the central portions of the chromosomes and the centromere to be located near one side of a low diversity region for all chromosomes except 9. However, there are many low diversity regions that are not centromeric (for example, the large region on chromosome 8). These low diversity regions are likely to represent regions in which B73 and Mo17 are identical by descent or regions with no structural variation in the maize species. These low diversity regions also exhibit very low levels of differential gene expression. Only three of the 196 genes from the MGSP filtered gene set that are located in the conserved chromosome 1 or chromosome 8 regions and queried by the Affymetrix 17K microarray exhibit evidence for differential expression in seedling, embryo or endosperm tissue from B73 and Mo17 [11]. The few cases of differential expression within sequence-conserved regions may reflect the action of trans-acting factors that are polymorphic between B73 and Mo17.

Mega-base sized B73-specific sequence

One visually striking feature in Figures 2 and Figure S11B) is the region on chromosome 6 (positions 42,211,131 – 44,706,565) that contains a cluster of B73>Mo17 probes. Closer inspection of this region indicates that the region of elevated structural variation is ~2.6 Mb (Figure 4A). The majority of probes in this region are either poorly conserved or not present among the Mo17 454 WGS sequences. This finding suggests that this 2.6 Mb sequence is present in the B73 genome but entirely absent from the Mo17 genome. Primer pairs designed based on the B73 sequence of this region were used to conduct PCR on B73 and Mo17 (Table S1). All 38 primer pairs amplified B73 but not Mo17. These primer pairs were also used to query for the presence of this 2.6 Mb segment in 22 other maize inbred lines. The data suggest that 16 of the inbreds contain this segment while the other 6 did not (Figure 4B). These inbreds seemed to contain (or lack) the entire segment as a haplotype block. It should be noted that both the CGH and PCR analyses suggest that all 2.6 Mb of sequence is missing in its entirety from the Mo17 genome and from the other six inbreds; neither it, nor components of it, are located at non-allelic positions.

Based on the filtered gene set from the MGSP there are 31 genes within the ~2.6 Mb B73-specific interval. RNA-seq experiments provide evidence for expression of 14/31 genes located within this interval in B73 shoot apical meristem tissue (Yi Jia, Kaz Ohtsu and Patrick S. Schnable, unpublished data) suggesting that many of the genes in this interval may be functional in B73. In addition, three of the genes

within this interval are detected by the Affymetrix 17K microarray. The expression of all three of these genes are detected in B73 but not in Mo17 [11]. Notably, intermediate levels of expression of these genes are also detected in B73xMo17 and Mo17xB73 hybrids.

The sequence of the B73-specific region does not exhibit similarity to the chloroplast or mitochondrial genomes. The genes present within this region do not shown synteny to any specific region of the rice genome but are found scattered across different rice chromosomes. Maize chromosome 6 is syntenic to rice chromosomes 5 and 6 [32]. However, there is a region near the centromere that does not show synteny with any rice chromosome and the 2.6 Mb segment is located within this region. Fine-scale analysis of the synteny in this region indicates that the distal sequence shows synteny to rice chromosome 5 while the sequence proximal to the B73-specific sequence is syntenous to rice chromosome 6. Hence, the B73-specific region is right at the point where the syntenic regions of maize chromosome 6 appear to have fused relative to rice chromosomes 5 and 6. The facts that many of these genes are expressed in maize and that many of the genes within this region are conserved in rice implies that the B73-region was likely selected in maize and have been deleted in the Mo17 haplotype.

Identification of copy number variants and genome content differences

In addition to this large region of genome variation on chromosome 6, we expected to identify numerous smaller copy-number variants (CNVs). As seen in the analysis of several well-annotated BACs, there are probes every ~400 bp in low-copy genomic DNA (Figure 1). CNVs can be discovered by assessing the behavior of adjacent probes to identify segments of DNA that give consistently altered signal from two genomes. The DNAcopy algorithm [34-35] was used to identify segments within the CGH dataset with a minimum length of 5 probes (Methods). This resulted in the identification of 53,589 segments that are within a single intra-BAC DNA sequence contig. The distribution of the average log2(Mo17/B73) values for each segment was well approximated by a normal mixture model with four components, each corresponding to a different class of segments (Figure 5). Because the component distributions overlap, there is uncertainty about the class membership of each segment. However, it is possible to calculate the probability that any particular segment belongs in a specific class based on the segment's average log2(Mo17/B73) value (Methods). Using such probabilities, each segment was classified into its most likely class. This is a relatively permissive approach towards identifying segments. We proceeded to further restrict the results to generate a subset of "stringent" segments that are at least 2,000 bp in length, include at least 10 probes, and, for B73>Mo17 and Mo17>B73 classes, exhibit at least a two-fold difference between average B73 and Mo17 signals (Table 3). The DNA segments from the different classes exhibit different distributions for segment length, probe number/segment and repetitive DNA content (Table 3; Figure S12).

The B73>Mo17 and Mo17>B73 segments represent DNA sequences that are variable between B73 and Mo17. B73>Mo17 DNA segments could be the result of CNV or differences in genomic content (PAV) between the two lines. In an attempt to distinguish between these two possibilities we determined the proportion of each segment that was non-repetitive and that could be aligned to Mo17 WGS sequence reads. If a large proportion of the segment was found in Mo17 then it is likely that the segment is a CNV, while segments that are missing from the Mo17 WGS likely represent PAVs. The distribution of Mo17 coverage was very different for B73>Mo17 segments compared to the other categories of segments (Figure S13). Over 50% of the B73>Mo17 DNA segments have less than 20% sequence coverage by Mo17 sequences. In the other classes, a majority of segments have >60% coverage by Mo17 WGS. We decided to split the B73>Mo17 segments into three subgroups. B73>Mo17_PAV (present-absent variation) segments exhibit less than 20% coverage by Mo17 WGS reads and are therefore likely present in the B73 genome and absent from the Mo17 genome. B73>Mo17_CNV segments exhibit at least 80% coverage in the Mo17 WGS sequences and are likely examples of CNV. The remaining B73>Mo17 sequences (20%-80% coverage) are denoted as B73>Mo17_Int. (Intermediate). As expected, the B73>Mo17_PAV segments have a greater signal difference between B73 and Mo17 than do the B73>Mo17_CNV segments (Table 3).

The segments from the middle two distributions in Figure 5 represent DNA sequences that are present at the same copy number in B73 and Mo17. The segments in the distribution with a peak at log2(Mo17/B73) = -0.43 were classified as B73≈Mo17_SNP while the segments in the distribution with a peak at log2(Mo17/B73=0 were simply classified as B73≈Mo17. An additional class, B73≈Mo17_Int. (intermediate), includes DNA sequences that couldn't be definitively classified in either one of these two distributions but had a cumulative estimated probability of membership in these two classes that was greater than 0.8 for these two classes. In general, the B73≈Mo17_SNP, B73≈Mo17_Int. B73≈Mo17 segments have similar characteristics (Table 3). The B73≈Mo17_SNP and B73≈Mo17_Int. DNA segments have slightly higher levels of signal in B73 than in Mo17 and are likely the result of the inclusion within the segment of several polymorphic probes. This is supported by the slightly higher rates of polymorphic probes or molecular markers within B73≈Mo17_SNP segments than B73≈Mo17 segments (Figure S12B; Table 4).

Characterization of CNVs and PAVs

The segment analysis identified a large number of DNA segments with variation in B73 and Mo17. There are 60 stringent Mo17>B73_CNV segments that are predicted to occur in more copies in Mo17 than in B73. There are 3,681 stringent B73>Mo17 segments including 356 segments that are CNVs and another 1,783 PAV segments that are putative examples of genome content variation.

Several different approaches were used to validate the structural variants identified in this study. The 1,783 stringent B73>Mo17_PAV segments are predicted to be present in the B73 genome but absent from the Mo17 genome. Over 20,000 primer pairs were designed (usually from B73 sequences) and used to perform amplification from B73 and Mo17 genomic DNA. The numbers of primer pairs within each class of segment were determined (Table 4). The proportion of primers that were polymorphic between B73 and Mo17 is much higher for B73>Mo17 segments. The fact that the majority of the B73>Mo17_PAV polymorphic primer pairs only amplify a band in B73 and not in Mo17 confirms that many of these segments are present in the B73 genome and missing in the Mo17 genome. The 356 B73>Mo17_CNV segments are predicted to occur in more copies in the B73 genome than in the Mo17 genome. BLAST searches of 100 stringent B73>Mo17_CNV sequences against the B73 genome finds that 92% are present in at least two copies. In comparison, only 7% of the B73≈Mo17 segments have multiple matches within the B73 RefGen_v1. A large proportion (55%) of the B73>Mo17_CNV segments includes tandem duplications. This suggests that there are a number of haplotype-specific tandem duplications. The 60 Mo17>B73 segments are predicted to occur in more copies in Mo17 than B73. qPCR was used to assess the copy number for 12 of the 60 Mo17>B73 segments in B73 and Mo17 genomic DNA (table S2). The increase in copy number in Mo17 relative to B73 was validated for 11 of the 12 segments tested. In three of the cases tested qPCR provides evidence for greater copy number differences than the CGH data, suggesting that the CGH copy number estimates may be conservative. In combination, these approaches provide validation for the three major classes of CNV segments.

The CGH analysis identified hundreds of candidate CNVs and thousands of PAVs. These sequences are spread throughout all ten of the maize chromosomes (Figure 6). The filtered set of 32,540 high quality gene annotations from the MGSP were compared to the stringent DNA segments (Table 5). Using fairly strict criteria (80% of gene sequence is contained within segment sequence) we find approximately 80% of the genes are located within the stringent segments. Almost 600 of these genes are located in the B73>Mo17 or Mo17>B73 segments, including 180 gene models located within B73>Mo17_PAV segments and another 50 gene models located within CNV segments. These genes within the PAV and CNV type segments include many different annotations and are not enriched for putative uncharacterized proteins. Interestingly, the proportion of genes with a paralog (defined as >85% identity and coverage) is higher for the B73>Mo17 segments (Table 5). A portion of the genes within these segments are queried by the existing 17K maize Affymetrix microarray. The proportion of genes that are differentially expressed (in B73 and Mo17 seedling tissue; data from [11]) is much higher for B73>Mo17 and Mo17>B73 segments than for B73≈Mo17 classes (Table 5). As expected, the B73>Mo17 segments are enriched for genes with higher expression in B73 than in Mo17 and the Mo17>B73 segments are enriched for genes with higher expression in Mo17.

Discussion

There is wide-spread appreciation for the high level of diversity within the maize species [8,12,31]. This diversity is critical for breeders to select for novel agronomic traits and is important for heterosis. The availability of a reference genome sequence for one inbred (B73; Schnable *et al.*, submitted) coupled to CGH technology, has provided the opportunity to study the structural variation present between two inbred lines, B73 and Mo17. The extensive structural variation between B73 and Mo17 includes copy number variants (CNV) and present-absent variation (PAV). However, despite the high level of variation genome-wide, there are many regions of the genome that have little or no variation. We will discuss the types of variation observed throughout the maize genome as well as the implications of this variation for phenotypic diversity and heterosis.

Low diversity regions in a highly polymorphic species

It is tempting to assume that all genomic regions are different in these two lines. However, by assessing the levels of variation along the B73 RefGen_v1 it quickly becomes obvious that this variation is not randomly distributed. We identified a number of large regions (>1 Mb) that have little or no variation. The fact that these regions co-localized with chromosomal regions that lack genetic markers that exhibit polymorphisms in the Intermated B73xMo17 (IBM) mapping population [14,36] demonstrates that these marker-depleted regions are simply due to low/no diversity between the two parents of the mapping population. In general, almost all of the large low diversity regions occur within regions of low recombination frequency. This could contribute to the inheritance of large chromosomal regions that are identical-by-descent. The centromeres of most chromosomes are located within or at one end of low-diversity regions.

Several groups have assessed molecular diversity in maize populations in studies designed to identify the targets of domestication and/or selection in maize [37-39]. We noticed that two of the genes known to have been targets of selection or domestication, y1 [40] and tb1 [41] are located within large low diversity regions. In addition, a 1 Mb region on chromosome 10 with evidence for a selective sweep [42] also occurs in a region with low levels of structural variation. The chromosomal positions of 42 genes with evidence for selective sweeps [37-38] were compared with the level of structural variation. Nearly half of these genes (20/42) were located within large blocks of low diversity identified in this study. This finding is consistent with the hypothesis that many putative selection genes identified by virtue of their limited sequence diversity were not actual targets of selection but simply happen to be located within large blocks of reduced variation, some of which may have arisen via selective sweeps.

Frequency of structural variation in maize genome

We have identified thousands of examples of structural variation between the B73 and Mo17 genomes. The term structural Variantion is used to describe both sequences that are present in both individuals but have different copy numbers (copy number variants; CNV) and sequences that are present in one individual but absent in another (presence-absence variant; PAV). The unknown order and orientation of intra-BAC DNA sequence contigs will potentially lead to an over-estimation of the number of structural variation events by splitting some events into two different segments. However, it will also lead to an under-estimation of the number of events due to the fact that some smaller structural variant events which will not have enough sequence or probes on either side of a contig border to be called. We found that the 3,789 stringent B73>Mo17 or Mo17>B73 structural variants represent a minimum of 2,056 unique events (must be separated from nearest structural variant by >200,000 bp).

Differing probe densities, algorithms and statistical criteria complicate comparisons of rates of structural variation among organisms [5]. However, it is quite clear that the maize genome has a high rate of structural variation compared with other species. In the human, rat, dog, mouse, macaque and chimpanzee genomes the average number of CNVs between two individuals is between 15 and 75 [43-48]. A high resolution study of eight human genomes [49] revealed only several hundred insertions and deletions, including CNV and PAV sequences, in the comparison of any two human genomes. In contrast, even after very stringent filtering we identified >3,700 CNV or PAV sequences that represent at least 2,000 events between these two maize genomes. This likely represents a very conservative estimate of the true number of CNV and PAV events in the maize genome. Previous analyses of BAC libraries have also found significant differences in genome content [25]. This high level of structural variation with frequent changes in genome content is reminiscent of the high level of variation observed in the E. coli genome [50-51], but is without precedent among higher eukaryotes. As the levels of structural variation are assessed for other species it will be interesting to determine whether maize has an unusually high level of variation relative to other plants and animals.

Mechanisms and impact of CNV

This study identified >400 putative CNVs between B73 and Mo17. A combination of genome homology searches and qPCR suggests that many of these sequences represent actual CNVs. There is evidence that these CNV can be the result of tandem duplications or duplications dispersed throughout the genome. There are a large number of tandem duplications in the maize genome [32]; some of these are NIPs [29] and 5% of these exhibit CNV in our data (Y. Kai, P. Schnable, unpublished data). This suggests that some of the differences in copy number between B73 and Mo17 are due to haplotype-specific tandem duplication events. Alternatively, some of the CNVs may be caused by duplication to non-

allelic positions. Differences in genome content at the bz1 locus [23] actually represent a CNV event in which both genotypes have copies of a sequence at a shared position and one of the genotypes has one or more additional copies at a non-shared location [26]. Many of these CNV are likely the result of Helitron-mediated movement of gene fragments [12,25-26]. There is also evidence that tandemly duplicated gene families such as zeins [52] or disease resistance genes [53] exhibit differences in copy number for different haplotypes, possibly as the result of recombination-based mechanisms as have been analyzed in detailed by Yandeau-Nelson *et al.* [54].

Previous studies have suggested a high rate of near identical paralogs (NIPs) in the maize genome [29]. It is likely that the formation (or removal) of NIPs may have been haplotype specific. There are 50 genes from the MGSP's filtered gene set within our stringently called CNVs. By relaxing our criteria only slightly, we identify CNV segments that contain 558 genes (Table s3). As most gene fragments were successfully removed from the MGSP's filtered gene set [32], these genes within CNV are likely to include functional genes. Because 12 of the 14 CNV genes that were assayed exhibit variable gene expression levels in B73 and Mo17 seedlings, these genic CNV may contribute to phenotypic diversity.

Many maize alleles that are known to be epigenetically regulated exhibit allelic variation for tandem repeats. There is allelic variation in the tandem duplication of coding regions at the p1, c2, and r1 loci that exhibit epigenetic regulation [54-56]. In addition, there is evidence for allelic variation in the copy number of a non-genic sequence ~100 kb upstream of the b1 gene that controls expression and paramutation [57]. It is possible that the high rates of CNV in both genes and other low-copy sequences contribute to high rates of expression variation and epigenetic regulation in maize.

Widespread genome content differences

In addition to the hundreds of CNV detected between B73 and Mo17 we also noted thousands of sequences that account for over 20 Mb of DNA that are present in the B73 genome and absent in the Mo17 genome. It is quite unexpected to find such a large number of sequences that are present within one haplotype of a species and missing from another. These include extreme examples such as the 2 Mb region on chromosome 6 as well as many smaller B73>Mo17_PAV sequences. Following the initial discovery of PAV sequences we sought to determine whether these PAVs included genes and to estimate the number of genes affected by PAV. Many PAV segments include genes contained within the MGSP filtered gene set. It is important to note that the MGSP filtered gene set was rigorously filtered to remove gene fragments and sequences with homology to transposable elements [32]. While it is possible that a subset of the PAV sequences may represent novel, uncharacterized transposable elements, it is clear that numerous genes are present in the PAV sequences. Several examples of putative genes that that are unlikely to represent transposable elements but that exhibit PAVs include GRMZM2G390498 (putative

superoxide dismutase), GRMZM2G066290 (putative pyruvate kinase), GRMZM2G139160 (C2H2 zinc finger protein) and GRMZM2G382393 (putative auxin efflux carrier).

It is, however, difficult to determine the exact number of genes affected by PAV because this number is strongly influenced by the stringency used to identify PAV sequences and by the criteria used to identify genes within the PAV sequences. Using quite strict criteria for identifying segments and genes within the segments, the PAVs include 180 genes from the MGSP's filtered gene set (Table S4) and the B73>Mo17_Int. segments include another 360 genes. These 180 and 360 genes all have at least 80% of the gene length included in the PAV sequence implying that these are full-length genes and not simply examples of PAV for gene fragments of the type reported by Morgante *et al.* (2005). A more permissive approach (Table S3) finds as many as 473 genes within the B73>Mo17_PAV and another 797 genes within the B73>Mo17_Int. seqments. The very conservative estimate of gene number, 180, or the more permissive estimate of gene number within PAV sequences, 1,270 (473 + 797), account for 0.5% or 4.0% of the genes within the MGSP's set of filtered genes, suggesting that PAV affects a significant portion of maize genes.

These present-absent sequences are spread throughout the B73 genome. These events differ in a significant way from those observed by Fu and Dooner [23] who detected copy number differences within small gene families. In contrast PAVs are low- or single-copy DNA sequences that occur in B73 and are not present anywhere in the Mo17 genome. Many these genes are expressed and as expected, the majority is expressed in B73 but not in Mo17. In addition, over 1/3 of the gene models within PAV sequences do not contain similar sequences located elsewhere in the B73 genome (Table S3). This suggests that the some examples of PAV (those with paralogs) may be functionally complemented by another gene but that a significant portion of the PAV sequences do not have a functional complement elsewhere in the maize genome.

The high level of PAV sequences between B73 and Mo17 may reflect ancient haplotype variation or more recent genomic rearrangements. We assessed the prevalence of 85 B73>Mo17_PAV segments in 22 other inbred lines (listed in Figure 4) using IDP primers [14]. Interestingly, all 85 of these segments are detected in at least two of the other inbred lines. The majority of these segments (53/85) are present in 30-70% of the other lines. The common presence/absence of these segments suggests that they often reflect relatively old events and not novel, inbred-specific, events.

While there is substantial phenotypic diversity between B73 and Mo17 even non-biologists quickly recognize both as corn plants. It is surprising that these inbreds can tolerate such a high level of genome content variation and still develop as "normal" corn plants. It is likely that deleterious PAVs have

been strongly selected against. Maize is normally an out-crossing species. Due to inbreeding depression, many of the first generation inbreds produced by breeders in the early part of the last century were drastically reduced in fitness, incapable of reproducing or even inviable. Those first-generation inbreeds that could be propagated were intercrossed to produce the second and subsequent generations of inbreds which comprise the commercial gene pool of maize. Hence, PAVs with strong effects on fitness would likely have been purged from the commercial gene pool. It will therefore be of great interest to explore the PAV content of landraces of maize that have not been subjected to the inbreeding bottleneck.

Potential impact of structural variation on phenotypic diversity and heterosis

The frequent CNV and PAV observed among maize inbreds may contribute to the high levels of phenotypic diversity and plasticity observed in maize. CNV and PAV can have significant contributions to phenotype. There is evidence that tandem duplications may be important for the evolution of traits such as disease resistance [58]. In addition, the variation in copy number may allow for the evolution of novel expression patterns. There is evidence that strong artificial selection on specific anthocyanin coloration patterns has often led to the formation of complex alleles with tandem duplications [56,59-60]. The presence of many CNV and PAV events provides opportunity for selection. As different structural variants are combined through breeding there is opportunity for novel trans-interactions and for formation of novel alleles through unequal crossing over. Long-term selection experiments (>100 years) have continued to make progress on quantitative traits [30] and it is possible that the genomic variation of maize provides source material to generate novel alleles.

The high levels of structural variation detected in this study and high levels of heterosis observed in certain hybrids of maize may be linked. Heterosis (the superior performance of a hybrid relative to its inbred parents) has pronounced and widespread effects on many traits. The high frequency of genome content differences suggests a large number of linked content differences. In this study we identify several thousand sequences that are present in B73 but missing in Mo17. If we assume that there are an equivalent number of sequences that are present in Mo17 but absent in B73 we would expect nearly 4,000 genome content differences distributed throughout the B73 and Mo17 genomes. The finding of single-copy, expressed PAVs among maize inbreds demonstrate that it will be important to obtain the genome sequences of a number of inbred lines to identify the full complement of genes present within the maize species. The large number of potential combinations of PAV sequences in hybrids also provides the opportunity for novel gene complements in hybrids relative to the parental lines. Previous analyses of gene expression in B73, Mo17 and the F1 hybrid identified a number of genes that are expressed in one parent but not the other [11]. Interestingly, all of these genes are expressed in the hybrid leading to a larger number of transcripts in hybrids than in the inbred parents. In addition, the combination of inbred-specific sequences in the hybrids provides opportunities for novel trans-interactions that would

not occur in either parent potentially leading to non-additive expression levels [10,61]. Hence, further explorations of genomic variation among maize lines may lead to opportunities to elucidate the mechanisms of heterosis.

Methods

Microarray design

An oligonucleotide microarray was designed by Roche NimbleGen to perform comparative genomic hybridization (CGH) of maize inbreds (Copies of this design may be acquired by ordering: 080418_zea_mays_B73_CGH_HX1). A set of 14,423 maize BACs (downloaded March 2008) was used to design isothermal probes, varying in length from 45 bp to 85 bp and with a target Tm of 76C at a fixed interval of 50 bp. Probe sequences were repeat-masked by calculating the average 14-mer frequency for each probe, based on a frequency table generated from the complete set of BAC sequences available as of that date, and removing probes with an average 14-mer frequency higher than 400. Probe uniqueness was determined by comparing each probe to B73 RefGen_v1, using SSAHA (http://www.sanger.ac.uk/Software/analysis/SSAHA/) with a step-size of 1, nmer-size of 12 and a minimum match length of 33 bp. Up to five insertions/deletions were allowed in each match. Probes with <=15 close matches in the genome were included in the array design. Median final probe spacing was 450 bp. It should be noted that this set of probes was designed to facilitate sequence capture [62]; if NimbleGen's CGH probe design criteria had been utilized it is likely that the choice of probes would have been slightly different.

Probe annotations

The sequences of CGH probes were aligned to the B73 RefGen_v1 (Schnable *et al.*, submitted); >90% of the probes (1,977,283/2,124,029) could be mapped with 100% identity and coverage (Figure S1). These include probes with a single match to the B73 RefGen_v1 as well as probes with multiple perfect matches (Figure S1). The remaining 146,746 probes (either imperfect matches or not found in the B73 RefGen_v1) had very low hybridization signals suggesting that they were likely artifacts created by using unfinished BAC for probe design and were therefore omitted from all subsequent analyses. The probes that could be mapped to the B73 RefGen_v1 were further annotated for repetitiveness, for gene annotation and for sequence conservation with Mo17. The "repetitiveness" of each of the probes was classified using a series of repeat filters. The ~3% (54,791) that match at least five locations in the B73

genome with >97% identity and coverage were designated as "multi-copy". The ~25% (530,314) that aligned to five or more genomic locations at reduced stringency (>90% identity and coverage) were designated as "crosshyb" probes. There were also 63,792 probes that match the ISU cereal repeat database (http://magi.plantgenomics.iastate.edu/) but did not meet the criteria for designation as multi-copy or icicle probes. Generally, the different classes of repetitive probes exhibit similar behavior and we will therefore refer to multi-copy, icicle and cereal repeat probes as "repetitive probes". The remaining 1,461,771 probes were designated as non-repetitive. The location of each probe relative to genic sequences was determined through comparisons to gene models provided by the MGSP and were assigned to the following classes: exon, exon-intron (crosses exon/intron border), intron, 5' (within 2,000 bp 5' of start site), 3' (within 2,000 bp 3' of the gene), intergenic (more than 2 kb from nearest gene). The conservation of sequences of individual probes to the Mo17 genome was classified via alignments to the 42,206,664 Mo17 WGS sequences provided by Daniel Rohskar from the DOE's Joint Genome Institute. Each probe was classified as perfect match (100% identity and coverage), highly conserved (>97% identity and coverage, not perfect match), conserved (97-90% identity and coverage, poorly conserved (75-90% identity and 70%-90% coverage not highly conserved) or no match (all other probe).

Microarray hybridizations

Total genomic DNA isolated from two-week-old etiolated seedlings of maize inbreds B73 and Mo17 were labeled and hybridized following the methods described in Selzer *et al.* [63] and Roche NimbleGen's CGH user's guide (see manufacture's User guide). In short, 1 ug of DNA was labeled using either 5' Cy3 or Cy5-labeled Random Nonamers (TriLink Biotechnologies). DNA was incubated for 2 hours at 37°C with 100 units (exo-) Klenow fragment (NEB) and dNTP mix (6 mM each in TE; Invitrogen). The labeled samples were then precipitated with NaCl and Isopropanol and then rehydrated in 25 µl of VWR H20. 34 µg of test and reference samples were combined in a 1.5 ml tube and dried down by SpeedVac. Samples were resuspended in 12.3 µl of H20 and 31.7 µl of Roche NimbleGen Hybridization Buffer (Roche NimbleGen Inc.) and incubated at 95°C. The combined and resuspended samples were then hybridized to the array for 60- 72 hours at 42°C degrees with mixing. Arrays were washed using Roche NimbleGen Wash Buffer System and dried using the NimbleGen Microarray Dryer (Roche NimbleGen, Inc). Arrays were scanned at 5 µm resolution using the GenePix4000B scanner (Axon Instruments). Data was extracted from scanned images using NimbleScan 2.4 extraction software (Roche NimbleGen, Inc.), which allows for automated grid alignment, extraction and generation of data files. In our experimental design we had seven replicates of B73 (one with Cy3 and six with Cy5) and seven replicates of Mo17 (six with Cy3 and one with Cy5). Images were processed and spatial normalization of data within the array was conducted according to Nimblegen's standard protocol. Due to the fact that our array was designed using B73 genomic sequence and the high rate of polymorphism between B73 and Mo17, the CGH data violated the assumption for the regularly used Q-spline normalization to make two channels of

hybridization intensities comparable (see supplemental text and figures for further information on normalization issues and solutions). We used a subset of 840,289 probes that were known to have identical sequence in B73 and Mo17 (based on B73 BAC sequences and Mo17 WGS data) as a control set to obtain a best-fitting cubic spline function [64], assuming that most of these probes should have the same hybridization intensities after normalization. The spline function was then globally applied for all probes to normalize the two channels. After within-chip normalization, linear model analyses using LIMMA [65-66] were conducted for the data from all microarrays. The linear model for each probe included effects for dyes and genotypes, and p-values were calculated to test for a signal difference between Mo17 and B73 genotypes as part of each linear model analysis. The p-values were converted to q-values which were used to control the false discovery rate as described by Storey and Tibshirani [67].

Segmentation analysis

A segmentation analysis was performed using DNAcopy [34-35]; with tune default parameter alpha=0.05, trim=0.05) to identify groups of probes that exhibit similar deviation from a log2(M/B) ratio of zero. These probes identify segments of DNA that have DNA sequence polymorphisms, altered copy number, or presence/absence in the two genotypes. The ordering and orientation of intra-BAC DNA sequence contigs, and therefore probe sequences, within a BAC has not been fully determined for most of the BACs. Consequently, the resulting segment predictions were split at intra-BAC DNA sequence contig boundaries following the DNAcopy analysis. All segments were assigned a unique identification and the average log2(M/B) for all probes within the segment was determined. The distribution of the average log2(M/B) across segments was modeled using a four-component normal mixture model [68]. The EM algorithm [69] was used to estimate the mixing proportion, the mean, and the variance associated with each of the four normal component densities, corresponding to four segment classes. Class membership probabilities for each segment were computed using the EM estimates. Each segment was then classified into one of the four classes if its most likely class was greater than 0.8. The segments were further filtered to remove all segments that contain fewer than 10 probes or 2000 bp of sequence to produce a set of stringent segments. The underlying sequence of these segments was obtained by parsing the segment data to produce a sequence that spanned the full segment. These segments were then further annotated by comparisons to repeats, gene predictions and Mo17 sequence. In order to classify a gene within a segment we required that 80% of the gene sequence be within the segment sequence.

Analyses of gene expression

Gene expression information was obtained from several different sources. The Affymetrix data was obtained from 11-day old seedlings (GEO: GSE8174; [11]) and cDNA microarray expression data was obtained 14-day old seedling tissue (GEO_ GSE3733; [10]). RNA-Seq data was obtained from B73 shoot apical meristem tissue (SAM) isolated as described by Ohtsu *et al.* [70]. A pool of RNA sample from L1 of

13 SAMs and a pool of RNA samples from L2 of 13 SAMs were extracted followed by RNA amplification and synthesis of double-stranded cDNA according to previous procedures [70]. The libraries were sequenced on the Solexa 1G Genome Analyzer at Canada's Michael Smith Genome Sciences Centre. Each library was sequenced using 2 lanes on a Solexa flow cell. The resulting Solexa reads were aligned to maize gene models (http://www.maizesequence.org) with the short read aligner NOVOALIGN (http://www.novocraft.com) using 32 bases. The low quality bases located at the end of reads were trimmed off by the program and only reads that mapped uniquely to the genome with a maximum of two mismatches including insertion/deletion (indel) across 32 bases were used for subsequent analyses. The reads uniquely mapped to genome were projected to gene models (release 4a.53).

qPCR validation of Mo17>B73 CNV

Primers were designed for 12 Mo17>B73_CNV segments (Table S2). 20ng of three biological replicates of genomic DNA isolated from B73 or Mo17 seedlings was amplified with Applied Biosystems SYBR Green 2X PCR Master Mix (Applied Biosystems) using an Applied Biosystems 7900HT Real-Time PCR System in a 20μl reaction volume. Two technical replicates were performed for each sample. The average cycle threshold (Ct) values were determined for the technical replicates. The relative copy number was determined by comparing the Ct value for the test primer set to three different genomic controls known to be present in one copy in each genome.

Acknowledgements

We thank Dan Rokhsar of the DOE's Joint Genome Institute for sharing the sequences of Mo17 WGS reads prior to publication, the maize genome sequence project for sharing the sequence of the B73 genome, the B73 RefGen_v1 genome and annotation data prior to publication.

References

01. Feuk L, Carson AR, Scherer SW (2006) Structural variation in the human genome. Nat Rev Genet 7: 85-97.

02. Sharp AJ, Hansen S, Selzer RR, Cheng Z, Regan R, *et al.* (2006) Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. Nat Genet 38: 1038-1042.

03. Beckmann JS, Estivill X, Antonarakis SE. (2007) Copy number variants and genetic traits: Closer to the resolution of phenotypic to genotypic variability. Nat Rev Genet 8: 639-646.

04. Cooper GM, Nickerson DA, Eichler EE (2007) Mutational and selective effects on copy-number variants in the human genome. Nat Genet 39: S22-9.

05. Scherer SW, Lee C, Birney E, Altshuler DM, Eichler EE, *et al.* (2007) Challenges and standards in integrating surveys of structural variation. Nat Genet 39: S7-15.

06. Sebat J (2007) Major changes in our DNA lead to major changes in our thinking. Nat Genet 39: S3-5.

07. Hurles ME, Dermitzakis ET, Tyler-Smith C (2008) The functional impact of structural variation in humans. Trends Genet 24: 238-245.

08. Flint-Garcia SA, Thuillet AC, Yu J, Pressoir G, Romero SM, *et al.* (2005) Maize association population: A high-resolution platform for quantitative trait locus dissection. Plant J 44: 1054-1064.

09. Buckler E, Gaut B, McMullen M (2006) Molecular and functional diversity of maize. Current Opinion in Plant Biology 9: 172-176.

10. Swanson-Wagner RA, Jia Y, DeCook R, Borsuk LA, Nettleton D, *et al.* (2006) All possible modes of gene action are observed in a global comparison of gene expression in a maize F1 hybrid and its inbred parents. Proc Natl Acad Sci U S A 103: 6805-6810.

11. Stupar RM, Springer NM. (2006) Cis-transcriptional variation in maize inbred lines B73 and Mo17 leads to additive expression patterns in the F1 hybrid. Genetics 173(4): 2199-2210.

12. Messing J, Dooner H (2006) Organization and variability of the maize genome. Current Opinion in Plant Biology 9: 157-163.

13. Vroh Bi I, McMullen MD, Sanchez-Villeda H, Schroeder S, Gardiner J, *et al.* (2005) Single nucleotide polymorphisms and insertion–deletions for genetic markers and anchoring the maize fingerprint contig physical map. Crop Sci 46: 12-21.

14. Fu Y, Wen TJ, Ronin YI, Chen HD, Guo L, *et al.* (2006) Genetic dissection of intermated recombinant inbred lines using a new genetic map of maize. Genetics 174: 1671-1683.

15. Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS (2007) SNP discovery via 454 transcriptome sequencing. Plant J 51: 910-918.

16. Tenaillon MI, Sawkins MC, Long AD, Gaut RL, Doebley JF, *et al.* (2001) Patterns of DNA sequence polymorphism along chromosome 1 of maize (zea mays ssp. mays L.). Proc Natl Acad Sci U S A 98: 9161-9166.

17. Ching A, Caldwell KS, Jung M, Dolan M, Smith OS, *et al.* (2002) SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genet 3: 19.

18. Brown WL (1949) Numbers and distribution of chromosome knobs in united states maize. Genetics 34: 524-536.

19. McClintock B, Kato A, Blumenschein A (1981) Chromosome constitution of races of maize. its significance in the interpretation of relationships between races and varieties in the americas. Mexico: Colegio de Postgraduados. 517 p.

20. Adawy SM, Stupar R, Jiang J (2004) Fluorescence in situ hybridization analysis reveals multiple loci of knob-associated DNA elements in one-knob and knobless maize lines. J Histochem Cytochem 52: 1113-1116.

21. Kato A, Lamb JC, Birchler JA (2004) Chromosome painting using repetitive DNA sequences as probes for somatic chromosome identification in maize. Proc Natl Acad Sci U S A 101: 13554-13559.

22. Laurie DA, Bennett MD (1985) Nuclear DNA content in the genera zea and sorghum. intergeneric, interspecific and intraspecific variation. Heredity 55: 307-313.

23. Fu H, Dooner HK (2002) Intraspecific violation of genetic colinearity and its implications in maize. Proc Natl Acad Sci U S A 99: 9573-9578.

24. Wang Q, Dooner HK (2006) Remarkable variation in maize genome structure inferred from haplotype diversity at the bz locus. Proc Natl Acad Sci U S A 103: 17644-17649.

25. Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, *et al.* (2005) Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. Nat Genet 37: 997-1002.

26. Lai J, Li Y, Messing J, Dooner HK (2005) Gene movement by helitron transposons contributes to the haplotype variability of maize. Proc Natl Acad Sci U S A 102: 9068-9073.

27. Yao H, Zhou Q, Li J, Smith H, Yandeau M, *et al.* (2002) Molecular characterization of meiotic recombination across the 140-kb multigenic a1-sh2 interval of maize. Proc Natl Acad Sci U S A 99: 6157-6162.

28. Brunner S, Fengler K, Morgante M, Tingey S, Rafalski A (2005) Evolution of DNA sequence nonhomologies among maize inbreds. Plant Cell 17: 343-360.

29. Emrich SJ, Barbazuk WB, Li L, Schnable PS (2007) Gene discovery and annotation using LCM-454 transcriptome sequencing. Genome Res 17: 69-73.

30. Moose SP, Dudley JW, Rocheford TR (2004) Maize selection passes the century mark: A unique resource for 21st century genomics. Trends Plant Sci 9: 358-364.

31. Springer NM, Stupar RM (2007) Allelic variation and heterosis in maize: How do two halves make more than a whole? Genome Res 17: 264-275.

32. Schnable PS, Ware D, Fulton RS, Stein JC,Wei F, *et al.* (2009) The B73 maize genome: complexity, diversity and dynamics. Science 326: doi:1126/ science.1178534.

33. Hsia AP, Wen TJ, Chen HD, Liu Z, Yandeau-Nelson MD, *et al.* (2005) Temperature gradient capillary electrophoresis (TGCE)--a tool for the high-throughput discovery and mapping of SNPs and IDPs. Theor Appl Genet 111: 218-225.

34. Olshen AB, Venkatraman ES, Lucito R, Wigler M (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 5: 557-572.

35. Venkatraman ES, Olshen AB (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics 23: 657-663.

36. Liu S, Yeh C-T, Ji T, Ying K, Wu H, *et al.* (2009) Mu Transposon Insertion Sites andMeiotic Recombination Events Co-localize with Epigenetic Marks for Open Chromatin across the Maize Genome. PloS Genet 5: e733. doi:10.1371/ journal.pgen.1000733.

37. Wright SI, Bi IV, Schroeder SG, Yamasaki M, Doebley JF, *et al.* (2005) The effects of artificial selection on the maize genome. Science 308: 1310-1314.

38. Yamasaki M, Tenaillon MI, Bi IV, Schroeder SG, Sanchez-Villeda H, *et al.* (2005) A large-scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement. Plant Cell 17: 2859-2872.

39. Doebley JF, Gaut BS, Smith BD (2006) The molecular genetics of crop domestication. Cell 127: 1309-1321.

40. Palaisa KA, Morgante M, Williams M, Rafalski A (2003) Contrasting effects of selection on sequence diversity and linkage disequilibrium at two phytoene synthase loci. Plant Cell 15: 1795-1806.

41. Doebley J, Stec A, Hubbard L (1997) The evolution of apical dominance in maize. Nature 386: 485-488.

42. Tian F, Stevens NM, Buckler ES,4th (2009) Tracking footprints of maize domestication and evidence for a massive selective sweep on chromosome 10. Proc Natl Acad Sci U S A 106: 9979-9986.

43. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, *et al.* (2006) Global variation in copy number in the human genome. Nature 444: 444-454.

44. Graubert TA, Cahan P, Edwin D, Selzer RR, Richmond TA, *et al.* (2007) A high-resolution map of segmental DNA copy number variation in the mouse genome. PLoS Genet 3: e3.

45. Chen WK, Swartz JD, Rush LJ, Alvarez CE (2009) Mapping DNA structural variation in dogs. Genome Res 19: 500-509.

46. Guryev V, Saar K, Adamovic T, Verheul M, van Heesch SA, *et al.* (2008) Distribution and functional impact of DNA copy number variation in the rat. Nat Genet 40: 538-545.

47. Lee AS, Gutierrez-Arcelus M, Perry GH, Vallender EJ, Johnson WE, *et al.* (2008) Analysis of copy number variation in the rhesus macaque genome identifies candidate loci for evolutionary and human disease studies. Hum Mol Genet 17: 1127-1136.

48. Perry GH, Yang F, Marques-Bonet T, Murphy C, Fitzgerald T, *et al.* (2008) Copy number variation and evolution in humans and chimpanzees. Genome Res 18(11): 1698-1710.

49. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, *et al.* (2008) Mapping and sequencing of structural variation from eight human genomes. Nature 453: 56-64.

50. Golubovsky M, Manton KG (2005) Genome organization and three kinds of heritable changes: General description and stochastic factors (a review). Front Biosci 10: 335-344.

51. Welch RA, Burland V, Plunkett G,3rd, Redford P, Roesch P, *et al.* (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic escherichia coli. Proc Natl Acad Sci U S A 99: 17020-17024.

52. Song R, Messing J (2003) Gene expression of a gene family in maize based on noncollinear haplotypes. Proc Natl Acad Sci U S A 100: 9055-9060.

53. Smith SM, Pryor AJ, Hulbert SH (2004) Allelic and haplotypic diversity at the rp1 rust resistance locus of maize. Genetics 167: 1939-1947.

54. Chopra S, Athma P, Li XG, Peterson T (1998) A maize myb homolog is encoded by a multicopy gene complex. Mol Gen Genet 260: 372-380.

54. Yandeau-Nelson MD, Xia YJ, Li J, Neuffer MG, Schnable PS (2006) Unequal sister chromatid and homolog recombination at a tandem duplication of the a1 locus in maize. Genetics 173: 2211-2226.

55. Della Vedova CB, Lorbiecke R, Kirsch H, Schulte MB, Scheets K, *et al.* (2005) The dominant inhibitory chalcone synthase allele C2-idf (inhibitor diffuse) from zea mays (L.) acts via an endogenous RNA silencing mechanism. Genetics 170: 1989-2002.

56. Walker EL, Robbins TP, Bureau TE, Kermicle J, Dellaporta SL (1995) Transposon-mediated chromosomal rearrangements and gene duplications in the formation of the maize R-r complex. EMBO J 14: 2350-2363.

57. Stam M, Belele C, Dorweiler JE, Chandler VL (2002) Differential chromatin structure within a tandem array 100 kb upstream of the maize b1 locus is associated with paramutation. Genes Dev 16: 1906-1918.

58. Cannon SB, Mitra A, Baumgarten A, Young ND, May G (2004) The roles of segmental and tandem gene duplication in the evolution of large gene families in arabidopsis thaliana. BMC Plant Biol 4: 10.

59. Pilu R, Piazza P, Petroni K, Ronchi A, Martin C, *et al.* (2003) Pl-Bol3, a complex allele of the anthocyanin regulatory Pl1 locus that arose in a naturally occurring maize population. Plant J 36: 510-521.

60. Ronchi A, Petroni K, Tonelli C (1995) The reduced expression of endogenous duplications (REED) in the maize R gene family is mediated by DNA methylation. EMBO J 14: 5318-5328.

61. Swanson-Wagner RA, DeCook R, Jia Y, Bancroft T, Ji T, Zhao X, *et al.* (2009) Paternal Dominance of Trans-eQTL Influences Gene Expression Patterns in Maize Hybrids. Science In press.

62. Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, *et al.* (2007) Direct selection of human genomic loci by microarray hybridization. Nat Methods 4: 903-905.

63. Selzer RR, Richmond TA, Pofahl NJ, Green RD, Eis PS, *et al.* (2005) Analysis of chromosome breakpoints in neuroblastoma at sub-kilobase resolution using fine-tiling oligonucleotide array CGH. Genes Chromosomes Cancer 44: 305-319.

64. Workman C, Jensen LJ, Jarmer H, Berka R, Gautier L, *et al.* (2002) A new non-linear normalization method for reducing variability in DNA microarray experiments. Genome Biol 3: research0048.

65. Smyth GK (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology 3.

66. Smyth GK (2005) Limma: Linear models for microarray data. In: Anonymous Bioinformatics and Computational Biology Solutions using R and Bioconductor New York: Springer. pp. 397-420.

67. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 100: 9440-9445.

68. Everitt BS, Hand DJ (1981) Finite mixture distributions. New York: Chapman & Hall. 143 p.

69. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society 39: 1-38.

70. Ohtsu K, Smith MB, Emrich SJ, Borsuk LA, Zhou R, *et al.* (2007) Global gene expression analysis of the shoot apical meristem of maize (zea mays L.). Plant J 52: 391-404.

71. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I (2004) VISTA: Computational tools for comparative genomics. Nucleic Acids Res 32: W273-9.

72. Wolfgruber TK, Sharma A, Schneider KL, Albert PS, Koo D, *et al.* (2009) Maize centromere structure and evolution: sequence analysis of centromeres 2.

Tables

Table 1. Probe classifications and enrichment in specific catergories

	All probes	Significant probes (q<0.0001)	B>M significant probes	M>B significant probes
	2,110,668	325,813	291,963	33,850
Probe classification by "repetitiveness"[a]				
Non-repeat	1,461,771 (69%)	278,390 (85%)	249,188 (85%)	29,202 (86%)
Total repetitive	630,586 (30%)	47,423 (15%)	42,775 (15%)	4,648 (14%)
Cereal repeat	226,706 (11%)	12,349 (4%)	11,586 (4%)	763 (2%)
Crosshyb	585,105 (28%)	40,907 (13%)	36,616 (13%)	4,291 (13%)
Multi-copy	54,791 (3%)	6,276 (2%)	5,497 (2%)	779 (2%)
Probe classification by Mo17 conservation annotation[b]				
Perfect match (100%)	871,664 (41%)	44,062 (14%)	22,586 (8%)	21,476 (63%)
Highly Conserved (>97%)	331,590 (16%)	42,992 (13%)	36,659 (13%)	6,333 (19%)
Conserved (>90%)	505,519 (24%)	106,996 (33%)	103,817 (35%)	3,179 (9%)
Poorly conserved (>75%)	182,570 (9%)	55,410 (17%)	53,929 (18%)	1,481 (4%)
No match in Mo17 (<75%)	219,325 (10%)	76,353 (23%)	74,972 (26%)	1,381 (4%)
Genic annotation				
Exon	98,886 (4.7%)	11,885 (3.6%)	9,803 (3.4%)	2,082 (6.2%)
Exon-intron	62,448 (3.0%)	8,447 (2.6%)	7,134 (2.4%)	1,313 (3.9%)
Intron	145,483 (6.9%)	24,047 (7.4%)	21,159 (7.2%)	2,888 (8.5%)
3' 2000bp	121,133(5.5%)	24,569(7.9%)	2,113(5.9%)	26,682(7.7%)
5' 2000bp	134,408(6.1%)	26,618(8.5%)	2,666(7.5%)	29,284(8.4%)
Intergene	1,650,875 (74.6%)	222,688 (71.4%)	24,717 (69.1%)	247,405 (71.1%)

[a]The repetitive nature of each probe was determined by comparing to the B73 reference genome. All probes that satisfy criteria (see methods for details) for cereal repeat, crosshyb or multi-copy were designated as repetitive.

[b]The probes were each classified based upon the most significant similarity to the Mo17 WGS sequence from JGI. The full definition for each category can be found in the methods section.

Table 2. Influence of polymorphisms on hybridization variation

# SNPs	# Probes	B>M significant probes	M>B significant probes	Average log2(M/B)
0	568	6 (1%)	6	0.000
1	180	13 (7%)	6	-0.196
2	95	18 (19%)	2	-0.400
3	67	17 (25%)	1	-0.552
4	54	14 (26%)	0	-0.733
5 or more	640	142 (22%)	9	-0.789
Total	1604	210 (13%)	24	

*Significant probes indicate a q value <0.0001 from the linear model

Table 3. Annotation and expressionof genes within the B73-specific interval of chromosome 6

Gene Model	RNA-Seq counts	Annotation
AC186585.4_FG002	0	Putative uncharacterized protein
GRMZM2G003461	155	Putative MURBZC
GRMZM2G009792	0	Drought inducible 22 kD protein
GRMZM2G014055	0	Thioredoxin H-type
GRMZM2G040833	0	Putative uncharacterized protein
GRMZM2G048791	1	Os07g0121800 protein
GRMZM2G050768	1	Bowman-Birk type trypsin inhibitor
GRMZM2G051686	0	Os07g0541400 protein
GRMZM2G066932	0	Putative uncharacterized protein
GRMZM2G066978	1	Putative uncharacterized protein
GRMZM2G076943	10	Putative uncharacterized protein
GRMZM2G082560	3	Plant viral-response family protein
GRMZM2G106513	0	Actin-like protein 3
GRMZM2G121024	61	ATFP3
GRMZM2G122520	127	Putative uncharacterized protein
GRMZM2G134020	26*	Putative uncharacterized protein
GRMZM2G150524	44	Putative uncharacterized protein
GRMZM2G156578	102	Putative uncharacterized protein
GRMZM2G161761	326	Putative uncharacterized protein
GRMZM2G320152	0	Putative uncharacterized protein
GRMZM2G338829	0	ZIM motif family protein
GRMZM2G348282	0	Putative uncharacterized protein
GRMZM2G372684	0	von Willebrand factor type A domain containing protein
GRMZM2G390512	0	NA
GRMZM2G403813	0	Os12g0149700 protein (Fragment)
GRMZM2G403828	1	Putative uncharacterized protein
GRMZM2G409627	0*	Putative TF-like protein
GRMZM2G410275	0	Putative uncharacterized protein
GRMZM2G433166	0	von Willebrand factor type A domain containing protein
GRMZM2G446078	0*	NA
GRMZM2G451679	60	Os12g0106500 protein (Nodulin-like family protein, expressed)

*Expression of these genes is detected by microarray analysis in B73 seedling, embryo and immature ea

Table 4. Characteristics of each category of DNA segment

All Segments

Type	n	Total # probes	Avg. log2(M/B)	Avg. length (bp)	Avg. probe #	Avg. % repetitive DNA	Avg. # genes	Avg gene density (per kb)	Avg % Mo17 coverage
B73>Mo17_PAV	4,222	52,618	-1.90	7,041	12.5	39.2%	0.6	0.090	7.6%
B73>Mo17_Int.	3,224	49,222	-1.56	8,859	15.3	57.5%	1.0	0.110	44.3%
B73>Mo17_CNV	729	10,312	-1.43	8,069	14.1	80.5%	0.6	0.076	96.9%
B73~Mo17_SNP	15,976	657,052	-0.43	36,253	41.1	67.6%	2.2	0.059	61.4%
B73~Mo17_Int.	9,573	391,681	-0.13	36,076	40.9	67.5%	2.2	0.060	72.3%
B73~Mo17	15,536	746,147	0.01	43,122	48.0	72.7%	2.1	0.048	78.2%
Mo17>B73_CNV	752	11,531	0.82	8,273	15.3	52.8%	1.4	0.167	84.4%
Unclassified	3,577	110,903	-0.60	26,756	31.0	65.7%	1.6	0.062	53.5%

Stringent segments (minimum of 10 probes, 2000bp and 2 fold change between B73 and Mo17 for B73>Mo17 and Mo17<B73 classes)

Type	n	Total # probes	Avg. log2(M/B)	Avg. length (bp)	Avg. probe #	Avg. % repetitive DNA	Avg. # genes	Avg gene density (per kb)	Avg % Mo17 coverage
B73>Mo17_PAV	1,783	33,773	-1.77	10,688	18.9	44.9%	0.9	0.081	9.3%
B73>Mo17_Int.	1,542	33,947	-1.46	12,698	22.0	55.2%	1.3	0.104	43.9%
B73>Mo17_CNV	356	6,846	-1.41	9,878	19.2	84.1%	0.8	0.078	96.7%
B73~Mo17_SNP	13,183	637,722	-0.43	41,514	48.4	67.1%	2.5	0.060	62.9%
B73~Mo17_Int.	7,526	377,693	-0.13	42,563	50.2	66.6%	2.6	0.061	75.1%
B73~Mo17	12,720	726,894	0.01	49,943	57.1	72.5%	2.4	0.048	80.6%
Mo17>B73_CNV	60	1,110	1.26	6,512	18.5	58.2%	1.1	0.171	94.1%
Unclassified	2,661	104,579	-0.64	32,619	39.3	65.5%	2.0	0.062	53.1%

Table 5. # Polymorphic markers within segments

	# Primers	% polymorphic	% PA[a]
B73>Mo17_PAV	203	75%	83%
B73>Mo17_Int.	288	50%	65%
B73>Mo17_CNV	36	53%	66%
B73~Mo17_SNP	7,946	28%	35%
B73~Mo17_Int.	4,484	17%	29%
B73~Mo17	5,756	6%	27%
Mo17>B73_CNV	9	0%	0%
Unclassified	891	31%	46%
Total	19,613	20%	37%

[a]The proportion of polymorphic primers that amplify a product in B73 but not in Mo17

Table 6. Genes in stringent segments

Type	n	Avg. length (bp)	# FGS genes[1]	Genes per segment	% of genes with paralog	# Affymetrix genes[2]	%DE genes[3]	%DE with B>M[4]
B73>Mo17_PAV	1,783	10,688	180	0.10	63.4%	36	69%	92%
B73>Mo17_Int.	1,542	12,698	360	0.23	61.6%	68	56%	92%
B73>Mo17_CNV	356	9,878	41	0.12	62.5%	7	71%	100%
B73~Mo17_SNP	13,183	41,514	10491	0.80	50.0%	3347	24%	49%
B73~Mo17_Int.	7,526	42,563	5748	0.76	49.3%	1806	18%	48%
B73~Mo17	12,720	49,943	7831	0.62	49.7%	2306	12%	33%
Mo17>B73_CNV	60	6,512	9	0.15	54.5%	7	100%	0%
Unclassified	2,661	32,619	1364	0.51	55.4%	361	31%	66%

[1]The FGS refers to the filtered gene set of high-quality annotations produced by the MGSP

[2]Number of genes on Affy platform that are expressed in B73 or Mo17 seedling tissue

[3]Percent of genes that are differentially expressed ($q<0.05$)

[4]Percent of differentially expressed genes that are expressed at higher levels in B73 than in Mo17

Figures

Figure 1. Significant hybridization differences are due to structural variation. (A) The B73 and Mo17 sequences for a portion of the 9009 locus (sequenced by Brunner *et al.*, 2005) were aligned using Vista (Frazer *et al.*, 2004) which displays the percent identity as a sliding window of 100bp (y-axis is 50% to 100% identity). The location of genes annotated by Brunner *et al.* 2005 (indicated by light blue sequences in the alignment) and repeat elements (the color coded track right above the alignments; pink indicates retrotransposons and orange indicates transposons) are shown above the VISTA alignment. The $\log_2$(Mo17 signal/B73 signal) is shown for each probe in this region. The red probes exhibit significantly different ($q<0.0001$) signal in B73 and Mo17. The blue line indicates a segment with altered hybridization that was identified using DNAcopy. There are also data tracks that display the repeat annotation and B73/Mo17 similarity for each probe. Note that these annotations are based on the genome-wide analysis, not detailed analyses of these regions. In (B) we present the annotation, alignment and CGH data for a portion of the 9008 locus (sequence and annotated by Brunner *et al.*, 2005).

Figure 2. Genomic distribution of $\log_2$(Mo17/B73) signals. The $\log_2$(Mo17/B73) hybridization intensities are plotted for each chromosome. Data points below the line indicate higher hybridization in B73 than in Mo17. The positions of the centromeres (Wolfgruber *et al.*, submitted) are indicated by black boxes. Note that there are chromosomal regions with high rates of variation (example near 42-44 MB on chromosome 6) and regions with low rates of variation (example from 140-160 MB on chromosome 8).

Figure 3. Identification of regions of low structural diversity. The proportion of probes that exhibit significantly higher hybridization to B73 genomic DNA than Mo17 ($q<0.0001$) was determined for a sliding window of 1Mb probes with increments of 0.33Mb. The approximate position of each centromere (from Wolfgruber *et al.*, submitted) is indicate by a red circle on each chromosome. The location of the *tb1* (Doebley *et al.*, 1997) and *y1* (Palaisa *et al.*, 2003) genes, which are known to have undergone selective sweeps, are indicated. The gene density (based on the filtered gene set from the MGSP) is shown below each chromosome. The gene density was determined based on the number of genes per Mb. The dark color indicates low gene density while the yellow color indicates higher gene density.

Figure 4. Characterization of 2Mb region on chromosome 6 that is present in B73 but missing in the Mo17 genome. (A) A 10Mb region on chromosome 6 is shown. The color coding for each probe indicates the level of conservation of the probe sequence to the Mo17 WGS sequence. The coordinates on the x-axis refer to base pair position within chromosome 6 of the B73 Refgen_v1. The 2.6Mb region from 42.2 to 44.8 is enriched for probes that are poorly conserved or have no match in the Mo17 sequence and the majority of these probes exhibit much higher signal in B73 than in Mo17. (B) The data from 38 primer pairs are shown. Blue indicates successful amplification for a particular inbred by primer combination while red indicates no amplification. The full set of 38 primer pairs (see supplemental table 1 for details) amplify products in B73 but not in Mo17.

Figure 5. Distribution of average log2(M/B) for DNA segments. The distribution of the average $\log_2$(M/B) across segments was modeled using a four-component normal mixture model (Everitt and Hand, 1981). The EM algorithm (Dempster, Laird, and Rubin, 1977) was used to estimate the mixing proportion, the mean, and the variance associated with each of the four normal component densities, corresponding to four segment classes (labeled with arrows). Class membership probabilities for each segment were computed using the EM estimates.

Figure 6. Distribution of CNV and PAV throughout the maize genome. The position and average $\log_2$(M/B) for each Mo17>B73_CNV, B73>Mo17_CNV, B73>Mo17_I and B73>Mo17_PA segment is plotted for all 10 maize chromosomes. The color coding indicates the type of segment. The positions of the centromeres (Wolfgruber *et al.*, submitted) are indicated by the black boxes.

Supplemental figure 1. Flow-chart detailing the mapping of probe sequences to the B73 RefGen_v1.
Supplemental Figure 2. Density plots of sample chip signal intensity before and after global q-spline normalization.
Supplemental Figure 3. Alterations of the distribution of log2(M/B) values following different normalization approaches.
Supplemental figure 4. Significant hybridization differences are due to structural variation.
Supplemental Figure 5. Distribution of hybridization values in B73 and Mo17.
Supplemental figure 6. Volcano and MA plots for classes of repetitive probes.
Supplemental figure 7. Repetitive probes rarely report variation in B73 and Mo17.
Supplemental figure 8. Annotation of probes that exhibit significant ($q<0.0001$) variation in hybridization to B73 and Mo17 genomic DNA.
Supplemental figure 9. Volcano and MA plots for differing levels of conservation in Mo17 sequence.
Supplemental figure 10. Rates of variation and chromosomal distribution of probes with different levels of B73-Mo17 sequence conservation.
Supplemental figure 11. Genomic regions of low (A) or high (B) levels of structural variation.
Supplemental figure 12. Annotation of probes that are within stringent segments that exhibit CNV or PAV.
Supplemental figure 13. Distribution of Mo17 coverage for DNA segments in each category.

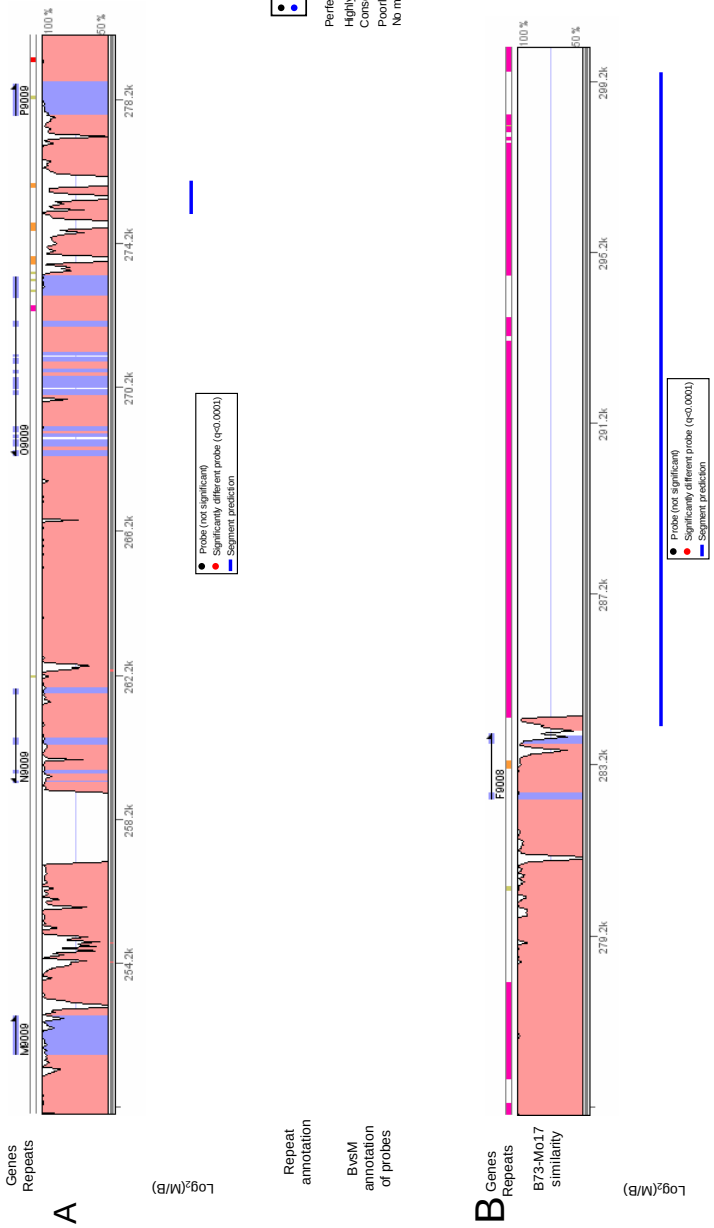


Figure 1. Significant hybridization differences are due to structural variation. The B73 and Mo17 sequences for two portions of the 9009 locus (sequenced by Brunner et al., 2005) were aligned using Vista (Frazer et al., 2004) which displays the percent identity as a sliding window of 100bp (y-axis is 50% to 100% identity). The location of genes annotated by Brunner et al. 2005 (indicated by light blue sequences in the alignment) and repeat elements (the color coded track right above the alignments; pink indicates retrotransposons and orange indicates transposons) are shown above the VISTA alignment. The log2(Mo17 signal/B73 signal) is shown for each probe in this region. The red probes exhibit significantly different (q<0.0001) signal in B73 and Mo17. The blue line indicates a segment with altered hybridization that was identified using DNAcopy. In (A) there are also data tracks that display the repeat annotation and B73/Mo17 similarity for each probe. Note that these annotations are based on the genome-wide analysis, not detailed analyses of these regions.

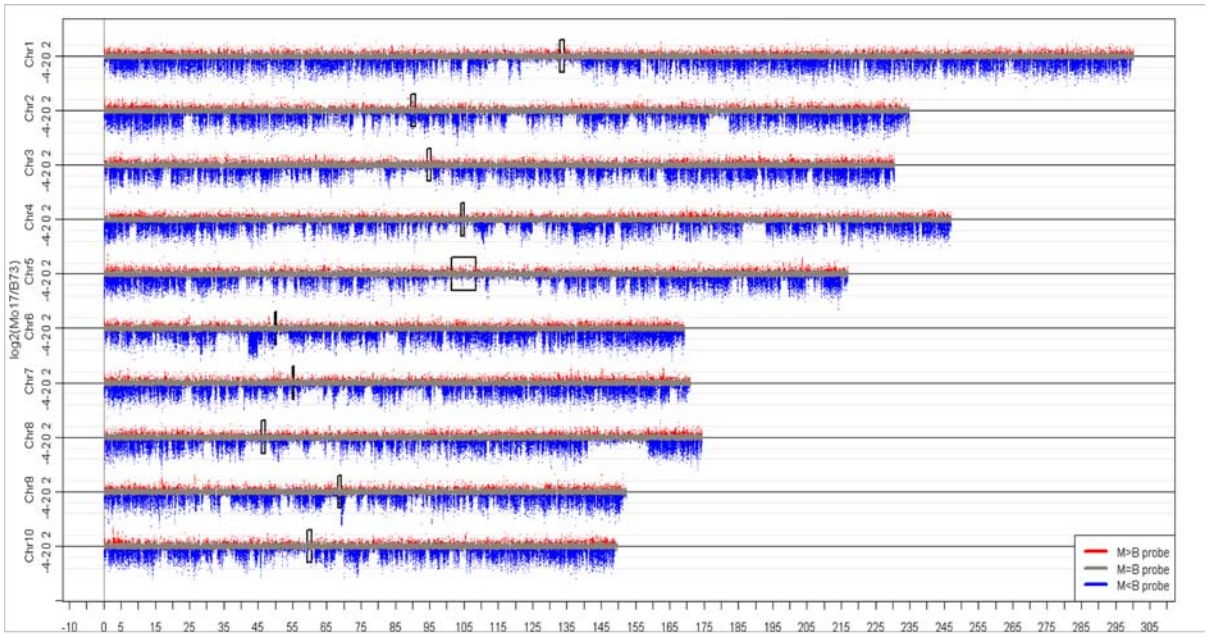


Figure 2. Genomic distribution of log2(Mo17/B73) signals. The log2(Mo17/B73) hybridization intensities are plotted for each chromosome. Data points below the line indicate higher hybridization in B73 than in Mo17. The positions of the centromere (Wolfgruber et al., submitted) are indicated by black boxes. Note that there are chromosomal regions with high rates of variation (example near 42-44 MB on chromosome 6) and regions with low rates of variation (example from 140-160 MB on chromosome 8).

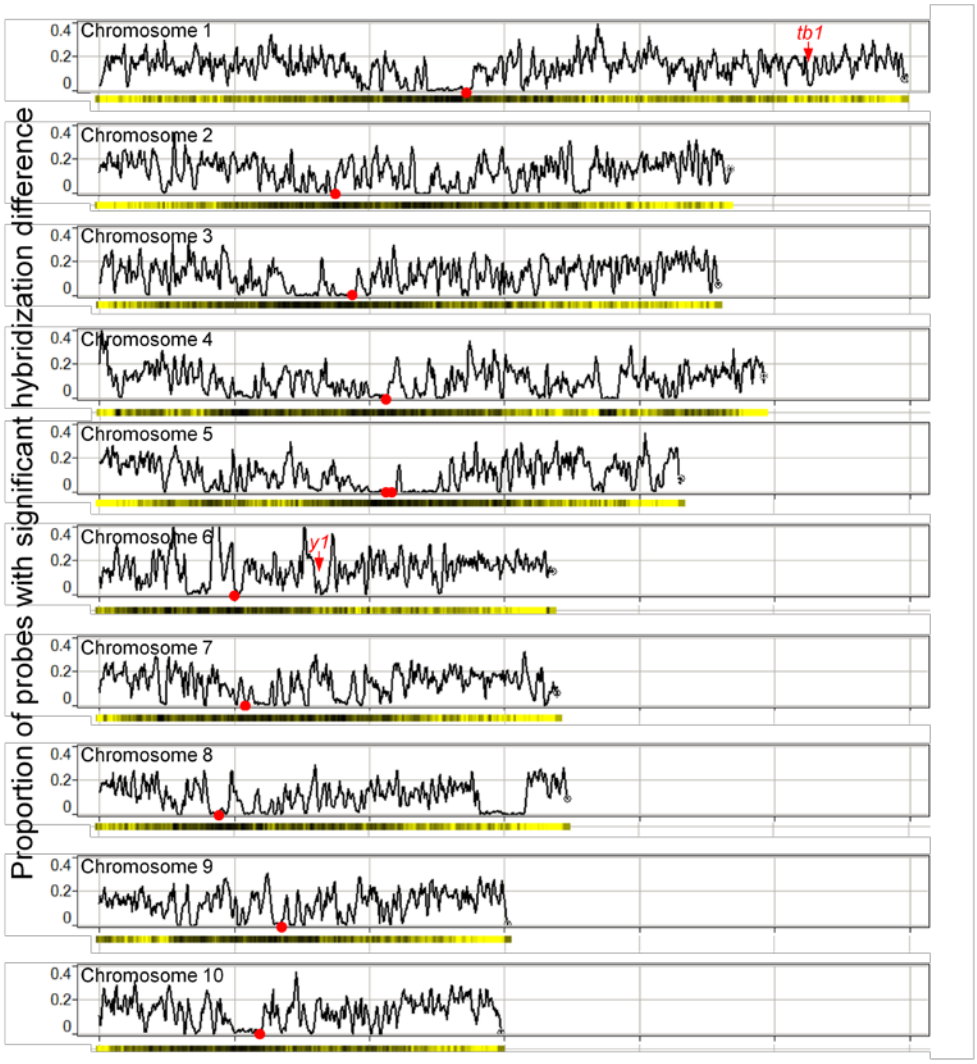


Figure 3. Identification of regions of low structural diversity. The proportion of probes that exhibit significantly higher hybridization to B73 genomic DNA (q<0.0001) was determined for a sliding window of 1Mb probes with increments of 0.33Mb. The approximate position of each centromere (from Wolfgruber et al., submitted) is indicate by a red circle on each chromosome. The location of the *tb1* and *y1* genes, which are known to have undergone selective sweeps, are indicated by asterisks. The gene density (based on the high-quality gene set from the MGSC) is shown below each chromosome. The gene density was determined based on the number of genes per Mb. The dark color indicates low gene diversity while the yellow color indicates higher gene density.

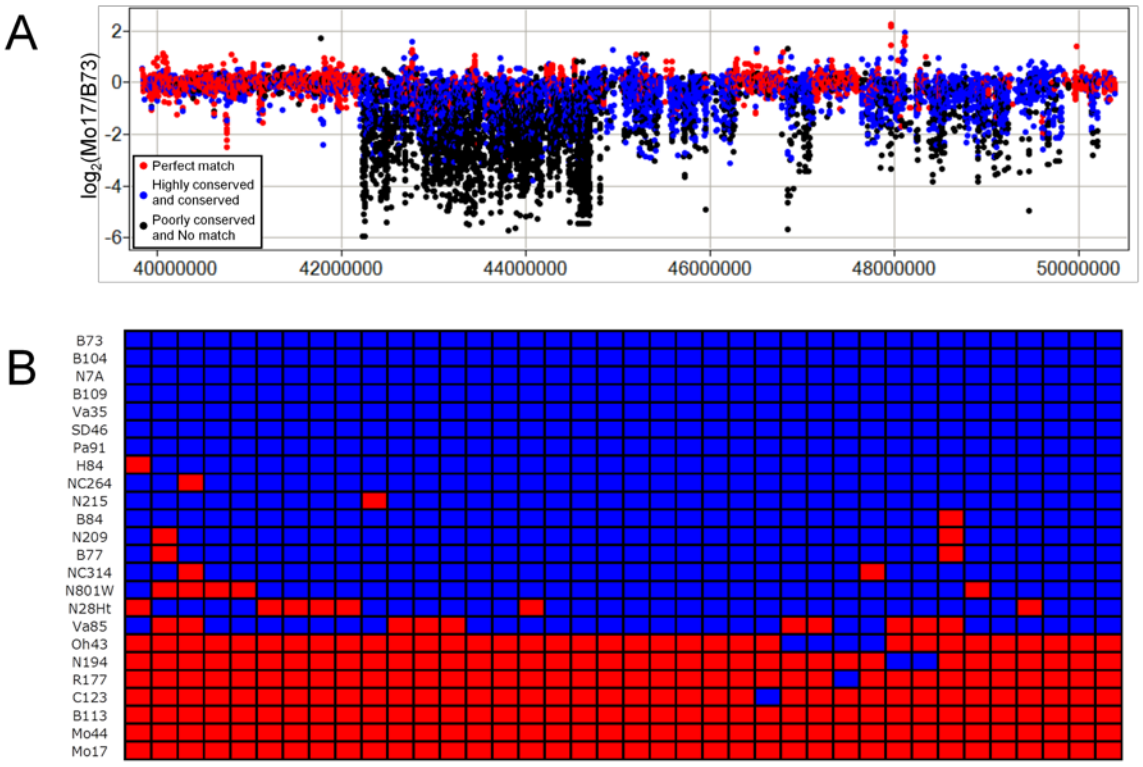


Figure 4. Characterization of 2Mb region on chromosome 6 that is present in B73 but missing in the Mo17 genome. (A) A 10Mb region on chromosome 6 is shown. The color coding for each probe indicates the level of conservation of the probe sequence to the Mo17 WGS sequence. The 2.6Mb region from 42.2 to 44.8 is enriched for probes that are poorly conserved or have no match in the Mo17 sequence and the majority of these probes exhibit much higher signal in B73 than in Mo17. (B) The data from 38 primer pairs is shown. Blue indicates successful amplification for a particular inbred by primer combination while red indicates no amplification. The full set of 38 primer pairs (see supplemental table 1 for details) amplify products in B73 but not in Mo17.

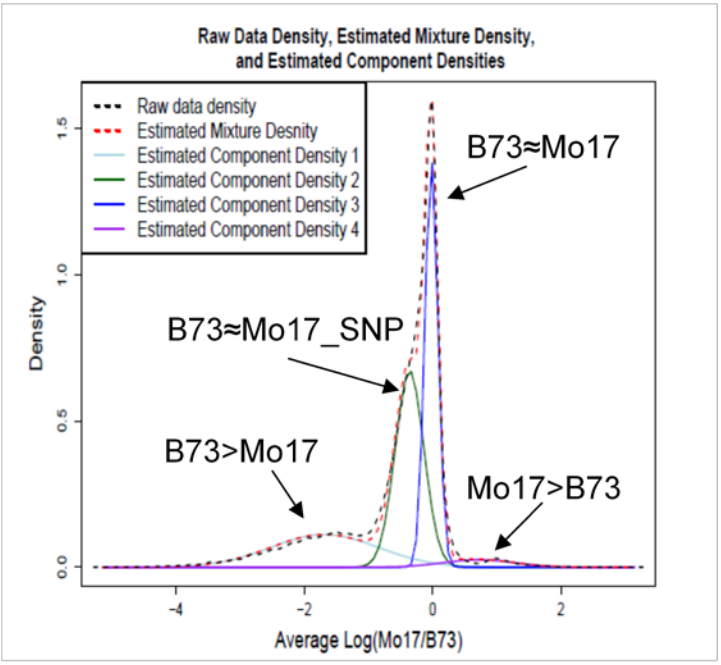


Figure 5. Distribution of average log2(M/B) for DNA segments. The distribution of the average log2(M/B) across segments was modeled using a four-component normal mixture model (Everitt and Hand, 1981). The EM algorithm (Dempster, Laird, and Rubin, 1977) was used to estimate the mixing proportion, the mean, and the variance associated with each of the four normal component densities, corresponding to four segment classes (labeled with arrows). Class membership probabilities for each segment were computed using the EM estimates.

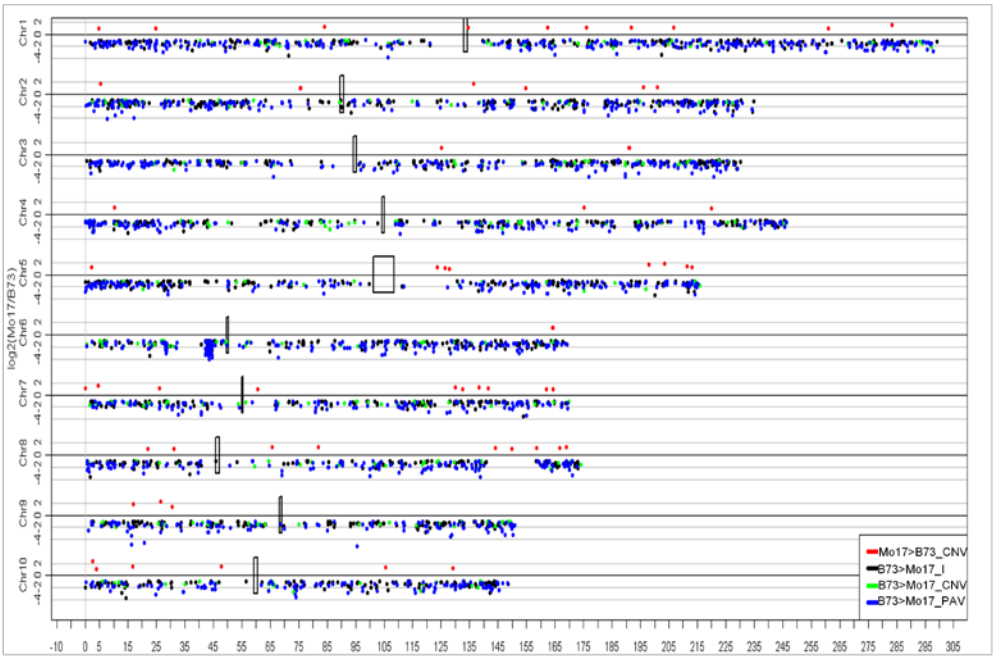


Figure 6. Distribution of CNV throughout the maize genome. The position and average log2(M/B) for each Mo17>B73_CNV, B73>Mo17_CNV, B73>Mo17_I and B73>Mo17_PA segment is plotted for all 10 maize chromosomes. The color coding indicates the type of segment. The position of the centromere (Wolfgruber et al., submitted) is indicated by the black box.

CHAPTER 3. ARRAY COMPARATIVE GENOMIC HYBRIDIZATION(ARRAY-CGH) BASED GENOTYPING FOR 2 LINES OF THE MAIZE INTER-MATED B73 X MO17 (IBM) MAPPING POPULATION.

Modified from a paper published in *PLoS one* 2010, 5(12): e14178

Yan Fu[1*], Nathan M. Springer[2*], Kai Ying[3,4*], Cheng-Ting Yeh[5], A. Leonardo Iniguez[6], Todd Richmond[6], Wei Wu[1], Brad Barbazuk[7], Dan Nettleton[8], Jeffrey A. Jeddeloh[6], Patrick S. Schnable[1,4,5**]

[1]Department of Agronomy, Iowa State University, Ames, Iowa, United States of America, [2]Department of Plant Biology, University of Minnesota, St Paul, Minnesota, United States of America, [3]Interdepartmental Genetics Graduate Program, Iowa State University, Ames, Iowa, United States of America, [4]Department of Genetics, Development and Cell Biology, Iowa State University, Ames, Iowa, United States of America, [5]Center for Plant Genomics, Iowa State University, Ames, Iowa, United States of America, [6]Roche NimbleGen, Inc., Madison, Wisconsin, United States of America, [7]Department of Biology and the Genetics Institute, University of Florida, Gainesville, Florida, United States of America, [8]Department of Statistics, Iowa State University, Ames, Iowa, United States of America

*These authors contributed equally
** Corresponding author: Roy J. Carver Co-Laboratory Iowa State University Ames, IA 50011-3650.
E-mail: schnable@iastate.edu

Abstract

To date, microarray-based genotyping of large, complex plant genomes has been complicated by the need to perform genome complexity reduction to obtain sufficiently strong hybridization signals. Genome complexity reduction techniques are, however, tedious and can introduce unwanted variables into genotyping assays. Here, we report a microarray-based genotyping technology for complex genomes (such as the 2.3 GB maize genome) that does not require genome complexity reduction prior to hybridization. Approximately 200,000 long oligonucleotide probes were identified as being polymorphic between the inbred parents of a mapping population and used to genotype two recombinant inbred lines. While multiple hybridization replicates provided ,97% accuracy, even a single replicate provided ,95% accuracy. Genotyping accuracy was further increased to .99% by utilizing information from adjacent probes. This microarray-based method provides a simple, high-density genotyping approach for large, complex genomes.

Introduction

The ability to rapidly determine genotypes at many loci in numerous individuals is critical to furthering our understanding of the inheritance of complex traits and for developing improved strategies for plant breeding. The use of molecular markers based on isozymes, RFLPs (restriction fragment length polymorphisms), SSRs (simple sequence repeats) and CAPS (cleaved amplified polymorphic sequences) genetic markers allowed for the construction of early genetic maps. However, these initial genotyping technologies were of relatively low throughput and required significant effort per data point. A number of technologies have been developed for high- throughput genotyping (reviewed by [1,2,3,4]). There are additional approaches that combine the use of microarrays and restriction digests such as diversity array technology (DArT) [5] and restriction site associated DNA (RAD) tags [6] to assay up to several thousand markers. These high-throughput approaches vary substantially in number of markers, amount of information required for development, accuracy, ease of application and data analysis. In particular, there are limitations on the application of some of these methods to species with large, complex genomes. The wide-spread availabilities of genomic and EST sequences in many species have led to the development of markers based on SNPs (single nucleotide polymorphisms). Several companies have developed high-throughput technologies that can genotype up to several hundred thousand SNPs in a single reaction [7,8]. Flibotte *et al.* [9] have reported the detection of SNPs in C. elegans (genome size 0.1 GB) using whole genome hybridization to arrays containing oligo probes designed based on the sequences of known SNPs (SNP-CGH). SNPs can be extremely valuable molecular markers but effort is required for the discovery and validation of SNPs, as well as for assay development. Alternative genotyping approaches that do not require prior knowledge of SNPs have also been developed. For example, RAD tags can be sequenced to discover and map SNPs [10]. Alternatively, SNPs and SFPs (single feature polymorphisms) can be detected by hybridizing genomic DNA or RNA to short oligonucleotide microarrays that contain short (~25 mer) oligonucleotides [11,12,13,14,15,16,17]. Longer oligonucleotide probes have also been used to detect SFPs caused by indel (insertion/deletion) polymorphisms [18,19,20]. Although this process is quite efficient in organisms with relatively small genomes, it has proven less successful for detecting polymorphisms in genomic DNA from organisms with larger genomes, such as maize (2.3 GB, [21]), due to a lack of sufficient signal strength. To obtain sufficient signal strengths in such species it is necessary to utilize RNA or reduced complexity (e.g., high Cot or methylation filteration) DNA. Unfortunately, these approaches introduce variables, such as expression level or filtration efficiency, that can complicate genotyping efforts. Previously we developed a custom long oligonucleotide microarray that yielded strong signals from whole genome hybridizations. This array was used to assess structural variation between the two maize inbreds B73 and Mo17 [22]. In that study ~200,000 probes were identified that exhibited highly differential and discriminatory hybridization signals between B73 and Mo17 genomic DNA. These hybridization differences were typically caused by the presence of multiple SNPs, small IDPs

(InDel Polymorphisms), and CNVs (copy number variants), including PAVs (presence absence variants), rather than by single SNPs [22]. To demonstrate the utility using these polymorphic probes as molecular markers for whole-genome genotyping in a large, complex genome, in this study two recombinant inbred lines (RILs) from the maize IBM mapping population [23] were randomly selected and analyzed via array comparative genomic hybridization (aCGH). In this report, we demonstrate the utility of using long oligonucleotide microarrays for high-throughput mapping in maize without the need to apply complexity reduction methods. The attractive features of this system are first, that it does not require prior knowledge of polymorphisms; second, that the genotyping results are highly accurate; third, high probe density allows for the fine-mapping of recombination breakpoints; and fourth, data analysis can be relatively simple.

Results

Identification of a large number of CGH-based polymorphisms

The first objective was to identify polymorphic probes that could be scored in the RILs. To identify such probes we used data derived from hybridization of the two parental genotypes (B73 and Mo17) to the microarray (all microarray data were deposited into GEO under Series# GSE16938). The microarray platform contained 1,262,421 probes that could each be unambiguously mapped to a single location (i.e., uniquely mapped) in the maize genome and that we therefore concluded are non-repetitive [22]. Following normalization and linear modeling (see [22] for full details), we identified 225,867 probes that exhibited significant differences in hybridization between B73 and Mo17 genomic DNA at a false discovery rate (FDR) cut-off of ,0.0001 (Table 1). The vast majority (91%) of these probes have higher hybridization signal intensities in B73 than in Mo17 and are referred to as B>M probes. Because the microarray was designed based on the reference B73 genomic sequence this observation is an expected consequence of ascertainment bias. B>M probes may have either of three possible characteristics. They can occur due to the existence of polymorphisms within the probe sequence between the two genotypes, due to the presence of more copies of the probe sequence in B73 than in Mo17, or due to a deletion of the probe sequence from the Mo17 genome. Because all probes were designed based on the B73 reference genome sequence, those probes that exhibit M>B hybridization ratios are expected to be present at higher copy number in the Mo17 genome than in the B73 genome.

An additional filter was applied to the derived probe list to cull for only the most utilitarian probes. This additional filter identified 173,122 probes that exhibited a minimum fold change of 2 (Table 1). This filter removed ,20% of the B>M probes and nearly 60% of the M>B probes.

The remaining 173,122 probes were annotated based on their genomic map positions relative to the B73 reference genome [21] and conservation in the Mo17 genome. Each probe sequence was compared with an ~5X whole-genome shotgun (WGS) sequence of Mo17 generated by the Joint Genome Institute using 454 sequencing technology (pre-publication access to these sequences was kindly provided by Dan Rokshar). Each probe was assigned a value of perfect match (100% identity and coverage), conserved (>90% coverage and identity) or ''no match'' (>90% coverage and/or identity). The majority (59%) of the B>M probes had no match in the collection of Mo17 WGS sequence reads, while only 1% had a perfect match (Table S1). Hence, as expected most (99%) of the B>M probe sequences are either absent from Mo17 or are polymorphic relative to B73. In contrast, the majority (51%) of the M>B probes had a perfect match in the collection of Mo17 WGS sequence reads, while only 13% did not have a match. Collectively, this set of polymorphic probes consisted of 173,122 probes, which included at least 12,000 probes for each of the 10 maize chromosomes (Table S2).

Assessment of data analysis and subsets of polymorphic probes

The potential of these polymorphic probes for genetic mapping was evaluated using two B73xMo17 recombinant inbred lines (IBM RILs; [23]) both of which we and others had previously genotyped using ~10,000 markers [24,25]. To enhance the utility of microarray-based genotyping we assessed the importance of hybridization replication, compared various methodologies for data analysis, and determined the effects of polymorphism types upon the accuracy of genotype determinations. Comparisons of the number of markers and accuracy of several different analytical approaches, including linear modeling of replicates, simple assessment of relative signals from a single replicate, and BAC- based genotyping were considered most germane. A visualization of the results obtained for chromosome 1 mapping was made for each of these approaches and is depicted in Figure 1.

The first analytical approach (Method I) involved the use of normalization methodology and subsequent estimation of the errors accounted for by dye and genotype effects upon the signal as determined via a linear model. This approach allowed for statistical contrasting of RIL vs. B73 and RIL vs. Mo17 at each probe using two hybridizations (See Methods for details). q-values were obtained for each of these contrasting comparisons and each probe was assigned to one of four classes in each RIL. Probes that were significantly different (q,0.05) from B73 but not from Mo17 in the RIL hybridizations were assigned a genotype of B (Class I) and probes that were different from Mo17 but not from B73 were

assigned a genotype of M (Class II). Some probes exhibited significant differences as compared to both parental lines (Class III) or were not significant in either of the two comparisons (Class IV). These later two classes may reflect non-polymorphic probes, residual heterozygosity or complex genome arrangements of gene families. Based on the broad genomic distribution of these probes (black dots in Figure 1) it is unlikely that residual heterozygosity is a major cause. Method I was able to assign genotypes for 93–95% of B>M probes in both RILs and was unaffected by the use of filtering based on a fold change (Table 1). However, substantially fewer of the M>B probes (74–86%) could be assigned genotypes (Table 1). Previously obtained genotyping results (from [25]) were used to validate the array-based genotyping calls (see Methods for details). As expected, consistency rates were substantially higher for the B>M probes than for the M>B probes and the use of the filtered probe set provided only a slight improvement in the validation (consistency) rate. While Method I provides robust results, it requires substantial bioinformatics expertise, as well as replication of hybridizations, factors that could discourage the broad adoption of this microarray-based genotyping platform.

Therefore, a more streamlined analytic method (Method II) was considered. This method employed a single hybridization and thus vastly reduced the complexity of the required computational analyses. In Method II, spatially normalized data from a single array were analyzed and hybridization contrasts were considered without applying statistical methods (See Methods for details). Our goal was to assess the relative loss of accuracy and information achieved using this relatively simple method of analysis and a single hybridization as compared to the robust Method I. The genotype for each probe was assigned by calculating the hybridization difference of the RIL and B73 relative to Mo17 and B73 [(RIL-B73)/(Mo17-B73)]. Probes with values near zero have hybridization intensities that are more similar to B73 than to Mo17, while values near 1 have hybridization intensities more similar to Mo17 than to B73. All probes with values less than 0.33 were assigned a genotype of B73 and probes with values greater than 0.66 were assigned a genotype of Mo17. The remaining probes were not classified (visualization provided in Figure 1). This approach assigned genotypes to slightly fewer probes than did the linear model (Method I) and had a slightly lower validation rate. Even so, this less complex analytic method still provided genotyping calls for >90% of the B>M probes and these calls were ~95% consistent with independently determined genotypes. Note that the filtered set of B>M probes provided substantially more benefit for Method II than for Method I. Consequently, rigorous filtering of probes is more critical when using a single hybridization (Method II) than when data from multiple hybridizations (Method I) are available. Based on a comparison of genotyping calls between two replicates (only two of the M0023 Cy3 replicates was used to enhance the consistency of analyses) the majority of the genotype assignments determined using B>M probes were consistent between pairs of replicates and only 2–4% of probes were called as different genotypes in the two replicates. As expected, the performance for the M>B probes was substantially lower for this approach. Only 40–60% of the M>B probes were consistently assigned to the

same genotype across independent replicates and the rate of inconsistent calls between the replicates was higher (Table 1).

Next, we investigated the utility of assigning genotypes to RILs based on a series of probes that were closely linked and that exhibited similar genotyping calls. The physical map of the maize genome is quite accurate at a resolution of single BACs [21]. However, the order and orientation of DNA sequences within BACs is often not known. This can lead to incorrect fine-scale arrangements in the order of probes in our genotyping data. Assigning each BAC a genotype in each RIL alleviates this problem. BAC-level genotyping is also expected to increase the accuracy of genotyping assignments because it allows for a genotyping assignment to be made using multiple probes located on the same BAC. In addition, doing so simplifies data visualization by reducing the number of data points. For this analysis we used only those B>M probes that had a minimum of a 2-fold change because this set exhibited the greatest accuracy in both Methods I and II. To be assigned a genotype a BAC had to have at least 5 probes that were assigned a genotype of B73 or Mo17 and these probes had to exhibit at least 80% genotype agreement within the BAC. Using this approach, genotypes could be assigned to over 95% the 8,497 BACs that contain at least 5 polymorphic probes (Method III, Table 2). The different methods of analysis were able to assign a genotype for slightly different sets of BACs (Table 2) but for BACs that were assigned genotypes with both methods there was 100% agreement of the genotyping calls made by the different approaches. By comparing the genotyping assignments with the genotyping data of Liu *et al.* [25] we could demonstrate >99% accuracy for each of these approaches. This approach of assigning a genotype for each BAC in each RIL allows for simple visualization of the genotyping calls (Figures 1 and 2).

We assessed the resolution of the genotyping data that was generated by CGH. While some recombination break-points can only be resolved within ~100 kb due to lack of polymorphic probes in the region of the recombination event other crossovers can be resolved at quite high resolution. Figure 3 provides five examples of highly resolved crossovers in the M0022 RIL. The exact location of these five crossovers could be identified within 2,450 to 6,042 base pairs. The ability to pinpoint the location of recombination events was influenced by the number of polymorphic probes within the region.

Discussion

This report documents the feasibility of genotyping complex genomes via a microarray-based method that does not require the use of methods to reduce genome complexity prior to hybridization. To do so, we used long oligonucleotides, which increased our ability to detect signal from genomic DNA and leveraged the abundance of frequent and widely distributed differences in DNA sequences between haplotypes as molecular markers. This approach is particularly valuable because it allows for mapping experiments even in the absence of any prior knowledge of polymorphisms between the parents of a mapping population, does not require extensive laboratory manipulation and can be performed using a single hybridization replicate. Additional experiments (data not shown) have demonstrated that this approach can be used to genotype individuals containing heterozygosity (such as F2 individuals) as well as homozygous RILs.

Comparison of linear model (Method I) and single array (Method II) analyses

Careful comparisons enabled us to determine the numbers of markers that could be genotyped, and their validation rates using replicated data analyzed using a linear model (Method I) as well as a more simplified analysis (Method II) conducted on non-replicated data. Replicates did provide slightly high numbers of markers that could be scored and yielded genotyping data that was validated at slightly higher rates. However, when probes were filtered to use only those that exhibited at least 2-fold change between the parental lines, the overall validation rates for Methods I and II were quite similar. We found that assigning genotypes to each BAC based on the occurrence of multiple polymorphic probes within the same BAC provided highly accurate genotyping scores. Indeed, this method resulted in nearly perfect validation rates (>99%). If it is desired to perform fine-scale mapping of recombination break-points to the highest possible resolution it would be possible to first conduct a BAC-based analysis to map recombination breakpoints to a BAC-level resolution. Subsequently, the analysis of individual probes within those BACs near the recombination breakpoint could further define the position of the recombination breakpoint.

Genotyping accuracy of different classes of probes

Because this method relies on long oligonucleotide probes many of the detected polymorphisms are likely to be structural variants. Because our probes were designed based upon the sequence of the B73 reference genome [21], each exhibits a perfect match to B73. Therefore, all probes with higher hybridization intensities in Mo17 than in B73 (M>B probes) are expected to represent sequences that are present in more copies in Mo17 than in B73 (CNVs). In contrast, the sequences detected by probes having higher hybridization intensities in B73 than in Mo17 (B>M probes) are likely to: 1) exhibit multiple sequence differences (SNPs and/or IDPs) between B73 and Mo17; 2) be absent from the Mo17 genome

(PAVs); or 3) exist in higher copy number in B73 than in Mo17 (CNV). Because only probes that had a single match to the B73 reference genome were used in this analysis, it is likely that most of the B>M probes are from the first two classes. Consistent with this view, a comparison of the B>M probes with a collection of Mo17 WGS sequences (~5X coverage) revealed that many do not have a 90% identical sequence. Although all probes used in this analysis are single copy in the B73 reference genome, it is, however, conceivable that some of the B>M probes match duplicated regions of the actual B73 genome that were either not sequenced or that were inadvertently collapsed into single copies during genome assembly. Hence, a small fraction of the B>M probe sequences may in fact exist at higher copy numbers in the B73 genome than in the Mo17 genome. Consistent with this possibility, a small fraction (1%) of the B>M probes had a perfect match to Mo17. This result is unexpected if indeed these probe sequences exist as single copy sequences in the B73 genome. The 13% of the M>B probes that did not have matches in the collection of Mo17 WGS sequence reads could be the result of inadequate sampling of the Mo17 genome. Although a small fraction of probes exhibit hybridization patterns that differ from expectations for various reasons, the overall mapping accuracy and resolution generated using this technology is high. It is noteworthy that the proportion of M>B probes that could be called and their validation rates were much lower than for the B>M probes. This likely reflects the fact that copy number variants (CNVs) could be in either the cis or trans configuration. If the multiple copies of a probe sequence are not closely linked in the Mo17 genome (i.e., in a trans configuration) they would be expected to segregate among RILs, potentially yielding novel hybridization signals (i.e., not similar to either of the parental signals).

Mapping strategy

In the experiments reported here we used an array that contained ~2.1 M probes. Roche NimbleGen also offers a customizable 12- plex suite of arrays. It contains 12 sets of the same 135,000 probes on a standard glass slide. Using the data obtained from a single dye-channel such a 12-plex array can be used to genotype 24 lines; these arrays have significant advantages from the perspectives of cost and efficiency. We have considered how best to modify our mapping strategy to accommodate the fact that each genotype will be analyzed with fewer probes (135,000 vs. 2.1 M). As a consequence of the high degree of sequence polymorphism in maize ,10% of the probes on our 2.1 M array (i.e., ,200,000) proved to be polymorphic between B73 and Mo17 even though our custom array was designed based on the B73 haplotype without reference to the sequence of the Mo17 haplotype. We expect quantitatively similar results would be obtained when comparing B73 to any other inbred that is not closely related to B73. Indeed, this was observed in comparisons of the inbreds Hp301 and Tx303 to B73 (unpublished observation). But importantly for the design of a mapping strategy, the same probes are not likely not be polymorphic in all comparisons or mapping populations.

We therefore recommend a two-step mapping strategy. In the first step a survey array containing

,2.1 M probes sampled from the low-copy, genic regions of the genome of interest will be used to identify probes that are informative in a given population via hybridizations to the parents of the mapping population. In the second step, based on the results of these hybridizations,,135,000 of the most informative polymorphic probes would be selected and used to construct 12-plex arrays for genotyping members of a mapping population.

We recommend for routine genotyping experiments using those probes that exhibit higher hybridization intensities from the reference genome (e.g., B>M probes) for initial mapping applications because they have higher validation rates. Subsequent analyses could use probes having higher hybridization intensities in the non-reference genome to estimate the relative rates of tandem and dispersed duplications.

Material & Methods

Plant materials

Genomic DNA was isolated from two-week-old seedlings of the inbreds B73 and Mo17 as well as from two IBM RILs: M0022 and M0023. According to previous genotyping results [24] the genomes of these two RILs are ,56% identical. 1 mg of DNA was labeled using either 59 Cy3 or Cy5-labeled Random Nonamers (TriLink Biotechnologies). DNA was incubated for 2 hours at 37uC with 100 units (exo-) Klenow fragment (NEB) and dNTP mix (6 mM each in TE; Invitrogen). Labeled samples were then precipitated with NaCl and isopropanol and rehydrated in 25 ml of VWR H20. 34 mg of test and reference samples were combined in a 1.5 ml tube and dried down using a SpeedVac. Samples were resuspended in 12.3 ml of H20 and 31.7 mlof NimbleGen Hybridization Buffer (Roche NimbleGen Inc.) and incubated at 95uC. The combined and resuspended samples were then hybridized to the array for 60–72 hours at 42uC degrees with mixing. Arrays were washed using NimbleGen Wash Buffer System and dried using a NimbleGen Microarray Dryer (Roche NimbleGen, Inc). Arrays were scanned at 5 mm resolution using a GenePix4000B scanner (Axon Instruments). Data were extracted from scanned images using NimbleScan 2.4 extraction software (Roche NimbleGen, Inc.), which allows for automated grid alignment, extraction and generation of data files. For this experiment, five hybridizations were performed and the samples hybridized to each array are as follows: Array 1 M0023 (Cy3)/B73 (Cy5); Array 2 M0022 (Cy3)/B73 (Cy5); Array 3 Mo17 (Cy3)/ M0023 (Cy5); Array 4 Mo17 (Cy3)/M0022 (Cy5); Array 5 M0023 (Cy3)/B73 (Cy5).

CGH data analyses

The probes were mapped to the B73 RefGen_v1 genome sequence [21] with 100% identity and coverage [22] and only probes with a single perfect match were used for this analysis. The integrated genetic and physical map of maize [25] was used to determine the physical location of each genetic marker on B73 RefGen_v1. The hybridization intensity of each mapped probe was estimated within each genotype using LIMMA[26] according to [22]. When applying q,0.0001 cutoff [27], a total of 225,867 probes that exhibited significantly different hybridization signals between B73 and Mo17 were deemed to be polymorphic. This set was further divided and filtered based on which genotype exhibited a higher signal and whether there was at least a 2-fold change in signal intensity between B73 and Mo17 (Table 1).

CGH-based Genotyping

Linear model. A linear model was used to calculate a q-value to estimate the false-discovery corrected probability that a particular probe was different from B73 or from Mo17. For each RIL, a probe was assigned a value of ‘‘B73’’ if it was significantly different from Mo17 ($q<0.05$) but not from B73 and was assigned a value of ‘‘Mo17’’ if it was significantly different from B73 ($q<0.05$) but not from Mo17. The probes that were significantly different from both or neither parental lines were not assigned a genotype.

Single array based model. A simple model was employed to assign genotype calls using a single replicate of data. The spatially normalized data extracted for each array using the NimbleScan software were imported into Excel. For each probe the value of [(RIL-B73)/(Mo17-B73)] was calculated. Cut-off values of 0.33 and 0.66 were arbitrarily selected for the purpose of this analysis. All probes having values of less than 0.33 was assigned a genotype of B73, while probes having values greater than 0.66 were assigned a genotype of Mo17. Probes with values between 0.33 and 0.66 were not classified. The values for different replicates were subsequently compared to determine the number of genotype assignments that were shared or conflicting for each of the hybridizations.

BAC based genotyping. Genotypes of each BAC were determined by comparing the calls for all polymorphic probes within a BAC. Genotypes were assigned to BACs have at least five polymorphic probes and only BACs with at least 80% agreement for the genotypes of all probes within the BAC that were classified as B73 or Mo17. The ‘‘consensus’’ genotyping assignments were assigned when a BAC was assigned the same genotype for each of the replicates for a RIL.

Validation of genotype assignments

The genotype scores for each of the two RILs were collected from a total of 10,143 markers [25] including IDP markers [24], TIDP markers [28], SNP markers [29] and other markers downloaded from

MaizeGDB (http://www.maizegdb.org). If one of these markers was located within 5,000 bp of a probe, the genotype obtained from this marker was treated as the ‘‘true’’ genotype for this probe. The proportions of the genotyping calls for probes that were supported by these other markers were then determined.

Acknowledgements

We thank An-Ping Hsia for useful scientific comments and assistance with manuscript preparation and Dan Rokhsar of the DOE’s Joint Genome Institute for sharing unpublished whole-genome shotgun sequences of the Mo17 genome.

Supporting Information

Table S1 Conservation of probe sequences in Mo17 whole genome shotgun sequence. Found at: doi:10.1371/journal.pone.0014178.s001 (0.04 MB DOC)

Table S2 Number of polymorphic probes per chromosome. Found at: doi:10.1371/journal.pone.0014178.s002 (0.04 MB DOC)

References

1. Kidgell C,Winzeler EA (2005) Elucidating genetic diversity with oligonucleotide arrays. Chromosome Res 13: 225–235.

2. Gilad Y, Borevitz J (2006) Using DNA microarrays to study natural variation. Curr Opin Genet Dev 16: 553–558.

3. Gunderson KL, Steemers FJ, Ren H, Ng P, Zhou L, *et al.* (2006) Whole-genome genotyping. Methods Enzymol 410: 359–376.

4. Gupta PK, Rustgi S, Mir RR (2008) Array-based high-throughput DNA markers for crop improvement. Heredity 101: 5–18.

5. Wenzl P, Carling J, Kudrna D, Jaccoud D, Huttner E, *et al.* (2004) Diversity Arrays Technology (DArT) for whole-genome profiling of barley. Proc Natl Acad Sci U S A 101: 9915–9920.

6. Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Res 17: 240–248.

7. Syvanen AC (2005) Toward genome-wide SNP genotyping. Nat Genet 37 Suppl: S5–10.

8. Fan JB, Chee MS, Gunderson KL (2006) Highly parallel genomic assays. Nat Rev Genet 7: 632–644.

9. Flibotte S, Edgley ML, Maydan J, Taylor J, Zapf R, *et al.* (2009) Rapid high resolution single nucleotide polymorphism-comparative genome hybridization mapping in Caenorhabditis elegans. Genetics 181: 33–37.

10. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3: e3376.

11. Winzeler EA, Richards DR, Conway AR, Goldstein AL, Kalman S, *et al.* (1998) Direct allelic variation scanning of the yeast genome. Science 281: 1194–1197.

12. Borevitz JO, Liang D, Plouffe D, Chang HS, Zhu T, *et al.* (2003) Large-scale identification of single-feature polymorphisms in complex genomes. Genome Res 13: 513–523.

13. Rostoks N, Borevitz JO, Hedley PE, Russell J, Mudie S, *et al.* (2005) Single- feature polymorphism discovery in the barley transcriptome. Genome Biol 6: R54.

14. West MA, van Leeuwen H, Kozik A, Kliebenstein DJ, Doerge RW, *et al.* (2006) High-density haplotyping with microarray-based expression and single feature polymorphism markers in Arabidopsis. Genome Res 16: 787–795.

15. Borevitz JO, Hazen SP, Michael TP, Morris GP, Baxter IR, *et al.* (2007) Genome-wide patterns of single-feature polymorphism in Arabidopsis thaliana. Proc Natl Acad Sci U S A 104: 12057–1206221.

16. Kumar R, Qiu J, Joshi T, Valliyodan B, Xu D, *et al.* (2007) Single feature polymorphism discovery in rice. PLoS One 2: e284.

17. Bernardo AN, Bradbury PJ, Ma H, Hu S, Bowden RL, *et al.* (2009) Discovery and mapping of single feature polymorphisms in wheat using Affymetrix arrays. BMC Genomics 10: 251.

18. Salathia N, Lee HN, Sangster TA, Morneau K, Landry CR, *et al.* (2007) Indel arrays: an affordable alternative for genotyping. Plant J 51: 727–737.

19. Edwards JD, Janda J, Sweeney MT, Gaikwad AB, Liu B, *et al.* (2008) Development and evaluation of a high-throughput, low-cost genotyping platform based on oligonucleotide microarrays in rice. Plant Methods 4: 13.

20. Singer T, Fan Y, Chang HS, Zhu T, Hazen SP, *et al.* (2006) A high-resolution map of Arabidopsis recombinant inbred lines by whole-genome exon array hybridization. PLoS Genet 2: e144.

21. Schnable PS, Ware D, Fulton RS, Stein JC,Wei F, *et al.* (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326: 1112–1115.

22. Springer NM, Ying K, Fu Y, Ji T, Yeh CT, *et al.* (2009) Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. PLoS Genet 5: e1000734.

23. Lee M, Sharopova N, Beavis WD, Grant D, Katt M, *et al.* (2002) Expanding the genetic map of maize with the intermated B73 x Mo17 (IBM) population. Plant Mol Biol 48: 453–461.

24. Fu Y, Wen TJ, Ronin YI, Chen HD, Guo L, *et al.* (2006) Genetic dissection of intermated recombinant inbred lines using a new genetic map of maize. Genetics 174: 1671–1683.

25. Liu S, Yeh CT, Ji T, Ying K, Wu H, *et al.* (2009) Mu transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome. PLoS Genet 5: e1000733.

26. Smyth G (2005) Limma: Linear models for microarray data. Bioinformatics and Computational Biology Solutions using R and Bioconductor. New York. pp 397–420.

27. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 100: 9440–9445.

28. Hsia AP, Wen TJ, Chen HD, Liu Z, Yandeau-Nelson MD, *et al.* (2005) Temperature gradient capillary electrophoresis (TGCE)–a tool for the high- throughput discovery and mapping of SNPs and IDPs. Theor Appl Genet 111: 218–225.

29. Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS (2007) SNP discovery via 454 transcriptome sequencing. Plant J 51: 910–918.

Tables

Table 1. Comparison of genotyping using different subsets of probes and multiple analysis methods.

	B>M (all[a])		B>M (2-FC[b])		M>B (all[a])		M>B (2-FC[b])	
Number probes	204934		164728		20933		8394	
RIL - M0022	% Calls[c]	% Consistent[d]	% Calls[c]	% Consistent[d]	% Calls[c]	% Consistent[d]	% Calls[c]	% Consistent[d]
Linear model calls (Method I)								
	95.2%	94.9%	95.3%	96.7%	85.9%	82.8%	78.4%	83.4%
Simple model calls (Method II)								
Replicate 1	93.1%	91.9%	93.4%	94.5%	74.2%	76.6%	73.6%	80.6%
Replicate 2	93.8%	93.6%	94.8%	96.0%	74.6%	79.5%	76.0%	81.8%
Consistency among replicates								
Both same call	84.0%	95.3%	87.3%	96.7%	44.2%	81.9%	49.9%	85.4%
Opposing calls[5]	4.2%		2.0%		11.5%		7.4%	
RIL -M0023	% Calls[c]	% Validated[d]	% Calls[c]	% Validated[d]	% Calls[c]	% Validated[d]	% Calls[c]	% Validated[d]
Linear model calls (Method I)								
	92.5%	96.7%	93.3%	97.1%	80.9%	86.4%	74.2%	87.7%
Simple model calls (Method II)								
Replicate 1	93.7%	93.9%	94.8%	96.3%	71.2%	80.5%	72.4%	84.7%
Replicate 2	96.6%	94.4%	97.4%	96.5%	73.2%	81.5%	72.7%	83.8%
Replicate 3	95.2%	94.2%	96.5%	96.5%	77.3%	82.9%	78.9%	85.4%
Consistency among replicates								
Same call in all replicates	90.9%	95.4%	94.1%	96.9%	58.0%	83.9%	60.9%	86.0%
Opposing calls [e]	1.8%		0.6%		3.3%		2.1%	

[a]The B>M or M>B (all) refers to all probes with a FDR<0.05 in a comparison of B73 and Mo17.
[b]The B>M or M>B (2-FC) probes refers to the subset of polymorphic probes that have a FDR<0.0001 and a minimum of 2-fold change between B73 and Mo17.
[c]The % calls is the percent of SFPs that could be assigned a genotype using an analysis method.
[d]The % consistent refers to the percentage of polymorphic probes that were assigned the same genotype in previously generated genotyping data.
[e]The opposing calls are those for which the same probe was assigned different genotypes in different replicates.
doi:10.1371/journal.pone.0014178.t001

Figures

Figure 1. Comparisons of analytical approaches for CGH-based mapping. Data for B>M probes located on chromosome 1 that exhibit at least a 2-fold change (n = 26,953) were plotted following the use of different data analysis approaches. The upper set of plots display data for the RIL M0022 while the lower panels show data for the RIL M0023. The plots on the left and right display probes from the entire chromosome 1 and a close-up view of a 20 Mb region of chromosome 1 (positions 200 Mb–220 MB), respectively. For each set of plots the first panel provides visualization genotyping calls and log2 (RIL/B73) ratios following normalization and analysis using a linear model of multiple replications (Method I). The second and third panels show the genotyping calls and log2(RIL/B73) ratios based on the analysis of a single replicate of data that was normalized using standard NimbleScan approaches (Method II). The final plot shows the genotyping calls for each BAC (n = 1,369) using data from replicate 1 (Method III).

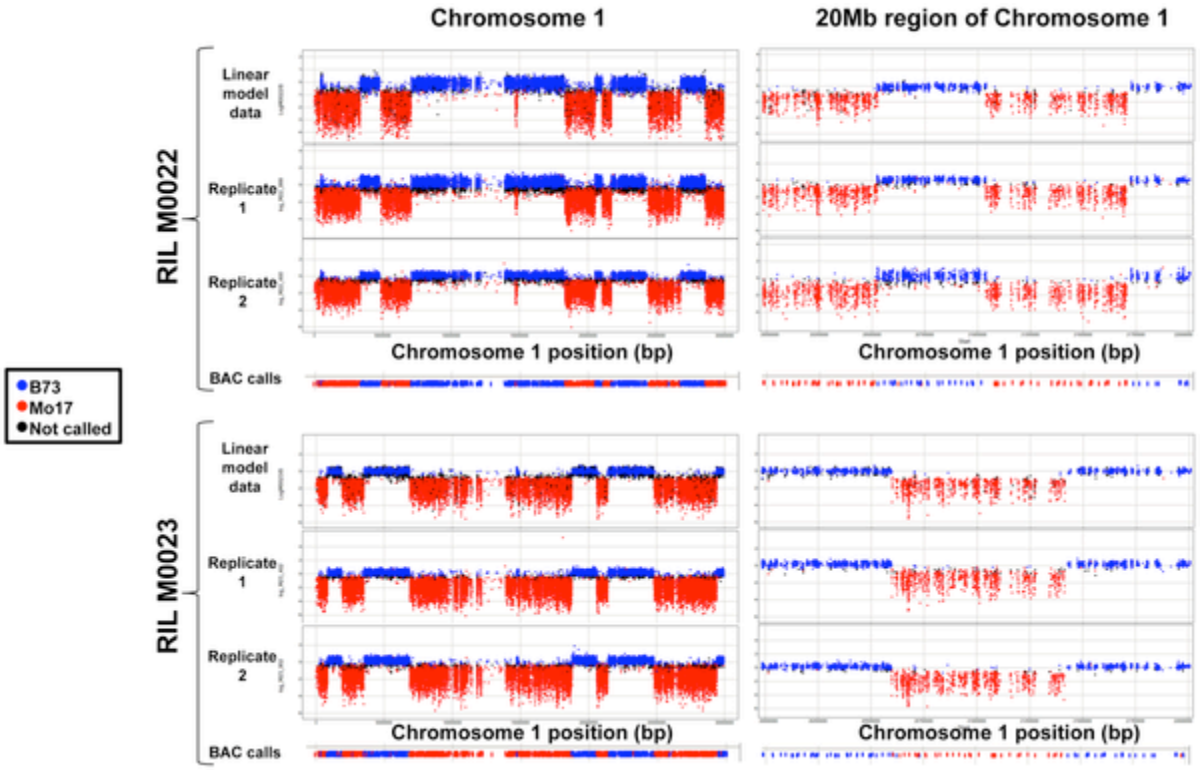


Figure 2. Visualization of whole genome genotypes of two RILs. The genotype of each BAC with at least 5 filtered B>M probes (n = 7,978 that were called in both lines) was determined and color coded (B73 – blue; Mo17 – red). BACs were then plotted according to their physical positions along a chromosome (x-axis) and by chromosome (y-axis). This visualization was created using a single replicate of data.

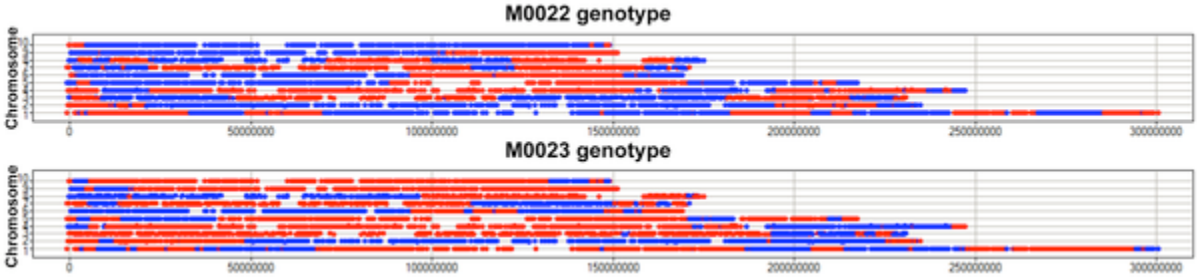
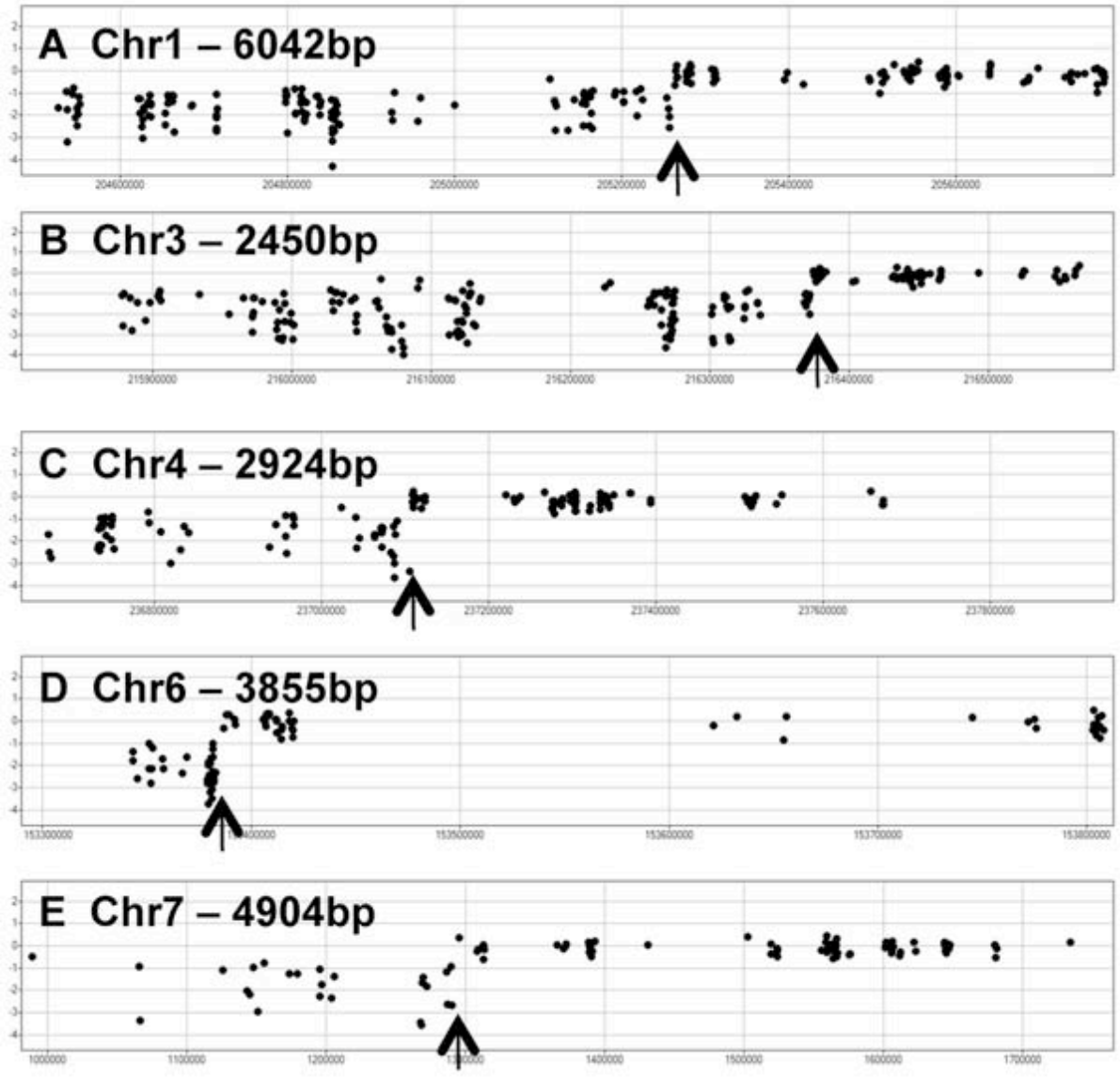


Figure 3. High-resolution of recombination break-points. Several plots show detailed views of the CGH mapping data near recombination events in the RIL M0022. The log2(M0022/B73) value is plotted along the y-axis for each of the B>M probes (q<0.0001 and fold-change>2) in five genomic regions. The arrowheads indicate the position of the recombination event and the label indicates the chromosome and the base pair resolution of the recombination event.



CHAPTER 4. CHANGES IN GENOME CONTENT GENERATED VIA SEGREGATION OF NON-ALLELIC HOMOLOGS

Modified from a paper published in *Plant J* 2012, in press

Sanzhen Liu[1], Kai Ying[2], Cheng-Ting Yeh[1], Jinliang Yang[2], Ruth Swanson-Wagner[2,3], Wei Wu[1], Todd Richmond[4], Daniel J. Gerhardt[5], Jinsheng Lai[6], Nathan Springer[7], Dan Nettleton[8], Jeffrey A. Jeddeloh[5], Patrick S. Schnable[1,*]

[1] Department of Agronomy, 2035 Roy J. Carver Co-Lab, Iowa State University, Ames, IA 50011;

[2] Interdepartmental Genetics, 2035 Roy J. Carver Co-Lab, Iowa State University, Ames, IA 50011;

[3] currently Monsanto, Amsterdam, the Netherlands;

[4] Research Informatics, Roche NimbleGen, 500 South Rosa Road, Madison, WI 53705;

[5] Development and Research, Roche NimbleGen, 500 South Rosa Road, Madison, WI 53705;

[6] State Key Lab of Agrobiotechnology, China Agriculture University, Beijing China;

[7] Department of Plant Biology, 250 Biological Science Center, St. Paul, MN 55108;

[8] Department of Statistics, 2115 Snedecor, Iowa State University, Ames, IA 50011

* Corresponding author:

Patrick S. Schnable,

2035B Roy J. Carver Co-Laboratory Iowa State University Ames, IA 50011-3650.

E-mail: schnable@iastate.edu

Abstract

A careful analysis of two maize recombinant inbred lines (RILs) relative to their inbred parents revealed the presence of several hundred apparently de novo copy number variants (CNVs). These

changes in genome content were validated via both PCR and whole exome-array capture-and-sequencing experiments. 185 genomic regions, which overlap with 38 high-confidence genes, exhibited apparently de novo copy number variation (CNV) in these two RILs and in many instances the same apparently de novo CNV events were observed in multiple RILs. Further analyses revealed that these recurrent apparently de novo CNVs were caused by segregation of single-copy homologous sequences that are located in non-allelic positions in the two parental inbreds. F1 individuals derived from these inbreds will be hemizygous for each of these non-allelic homologs but RIL genotypes will contain these sequences at zero, one or two genomic loci. Hence, the segregation of non-allelic homologs may contribute to transgressive segregation. Indeed, statistical associations between phenotypic QTL and genomic losses were observed for two of 14 tested pairs of non-allelic homologs.

Introduction

Genomes were once thought to be relatively stable structures having conserved gene order and gene content among individuals within a species. More recently, it has been demonstrated that the genomes of two individuals of the same species can exhibit substantial structural variation. For example, copy number variation (CNV), which refers to the number of copies of a particular sequence in different individuals of the same species, has been observed in many species (Chen *et al.* 2009, Conrad *et al.* 2010, Daines *et al.* 2009, DeBolt 2010, Fadista *et al.* 2008, Fadista *et al.* 2010, Hurwitz *et al.* 2010, Iafrate *et al.* 2004, Sebat *et al.* 2004, She *et al.* 2008, Springer *et al.* 2009). Geneticists are still exploring the ramifications of this intra-specific structural variation.

The maize lineage experienced a whole genome duplication ~5-12 million years ago (MYA), followed by gene fractionation which removed one member from some pairs of duplicated genes (Langham *et al.* 2004, Woodhouse *et al.* 2010). In addition transposons and perhaps other mechanisms have duplicated and transposed genomic sequences (including genes) throughout the genome (Bennetzen 2000, Jiang *et al.* 2004, Lai *et al.* 2005, Zhang and Peterson 2005). In combination these processes have resulted in a large and complex genome. Maize exhibits high levels of intraspecific variation (Lai *et al.* 2010, Schnable *et al.* 2009), including thousands of CNVs and genic presence-absence variants (PAVs) (Belo *et al.* 2010, Springer, *et al.* 2009, Swanson-Wagner *et al.* 2010). Maize is an excellent model for the study of quantitative trait variation. It exhibits prevalent transgressive segregation such that the recombinant offspring of two inbred varieties can exhibit phenotypes outside of the parental range.

In this report we document the existence of numerous non-allelic homologs in maize. These non-

allelic homologs represent single copy sequences that are present at different chromosomal locations in different individuals. A detailed analysis of CNVs was undertaken using several RILs (recombinant inbred lines) relative to their inbred parents. We detected more examples of apparently de novo CNV than expected and noted that several of these apparently de novo CNV were found in multiple RILs. Our investigations of these instances of recurrent apparently de novo CNV (Brunetti-Pierri *et al.* 2008, Fernandez *et al.* 2010, Neill *et al.* 2011, Shinawi *et al.* 2010) revealed that they were in at least most instances the result of segregation of non-allelic homologs (SNH), which generated RILs that completely lack or have extra copies of a given sequence. Finally, we provide evidence that these changes in sequence content can contribute to phenotypic variation.

Results

Chromosomal segments that exhibit non-parental-signals

Array-based comparative genomic hybridization (aCGH) experiments were conducted using genomic DNAs from two maize inbred lines, B73 and Mo17 (Springer, *et al.* 2009), and two of the IBM (Intermated B73 x Mo17) RILs (Lee *et al.* 2002): M0022 and M0023 (Fu *et al.* 2010) derived from these parental inbreds. A careful analysis of the aCGH signals of single-copy (i.e., those mapped to single genomic positions in the B73 genome) probes revealed that 0.06% (1,086/1,780,475 in RIL M0022) and 0.13% (2,338/1,780,475 in RIL M0023) of probes yielded signals in the RILs that were statistically different from those of both parents (Figure 1, Figure S1, S2, S3, Table S1). We defined these probes as putative de novo CNV probes. Many of these probes exhibiting unique levels of hybridization signal in the RILs relative to the parental genotypes can be grouped into "segments" using DNAcopy software (Methods). There are 67 chromosomal segments in RIL M0022, and 130 in RIL M0023 (Table 1 and Table S2) that represent putative de novo CNV events in the RILs. These include both gain and loss events, indicating copy number gain and loss, respectively. The average length of these CNV segments is 2.8 kb with the longest being 53.4 kb (Figure S1). The majority (69%) of the 197 chromosomal segments do not exhibit substantial differences in aCGH signal between B73 and Mo17. Hence, these aCGH results indicate that although the corresponding probes and chromosomal segments do not exhibit CNV between the inbred parents, these sequences do exhibit CNV in the RILs. These putative de novo CNV comprise 0.009% and 0.012% of the M0022 and M0023 genomes, respectively.

To confirm and extend these results a NimbleGen whole exome-array was used to capture and

sequence the genic portions of the chromosomal segments that exhibit non-parental hybridization levels (Methods). We separately (N=4) captured genomic DNAs from the parental inbreds (B73 and Mo17) and the two RILs (M0022 and M0023) using a previously published protocol (Haun *et al.* 2010). 32-37 million 40-bp paired-end Illumina reads generated from each capture were aligned to the B73 reference genome (GenBank accession no. SRA036595). Reads that uniquely mapped to the 197 chromosomal segments identified via aCGH were counted for all four genotypes. Figure 2 compares the aCGH and exome-Seq results of two chromosomal segments in both RILs. For subsequent analyses we focused on the 61 segments for which at least 30 reads were obtained from the B73 capture. Of these, 54 and 7 segments had exhibited aCGH signal losses (copy losses) and gains (copy gains) in the RILs relative to B73, respectively (Figure S4). The aCGH and exome-Seq count data for these segments are highly correlated (correlation=0.62). In the vast majority (45/54) signal loss segments, at least 80% fewer reads were obtained from the affected RIL than from B73. Notably, no reads were obtained from the affected RIL in 24 (of 54) chromosomal segments that exhibited signal loss in the aCGH experiments, demonstrating that these segments are completely absent from the RIL genome. Consistently, more reads were recovered from the RILs than from B73 for all 7 segments that exhibited signal gains in the RILs relative to B73, indicating that the RIL genomes contain more copies of these segments than do the parental genomes (Figure S4). Similar results were observed in comparisons to Mo17 (Figure S4). These exome-Seq results confirm the existence of apparently de novo CNV for many of the chromosomal segments in the RIL genotypes.

Apparently de novo CNV are the result of segregation of non-allelic homologs (SNH)

There are a number of potential mechanisms for de novo CNV formation including non-allelic homologous recombination (NAHR), rearrangements in the absence of extended sequence similarity associated with DNA repair by non-homologous end-joining (NHEJ) or with microhomology-mediated break-induced replication (MMBIR), contraction or expansions of variable number tandem repeats (VNTRs) and mobile element insertions (MEI) (Mills *et al.* 2011). Alternatively, apparently de novo CNV can be formed by segregation of single copy sequences that are located in non-allelic positions. If two parental lines both contain a single copy of a sequence that is located at unlinked genomic positions, then the F1 will be hemizygous for each of these copies and meiotic segregation will generate F2 (or RIL) genotypes with zero, one or two copies of the sequence (Lu *et al.* 2012). The relatively high rate of apparently de novo CNV observed in the RILs suggested that segregation of non-allelic homologs (SNH) might be responsible.

Several lines of evidence support this hypothesis. First, we noted that the locations of copy number gains and losses in the RILs exhibit dependence on the parental origins of the chromosomal segments containing these gains or losses. Those associated with significant signal losses in the RILs are

embedded within Mo17-derived chromosomal regions, while 19/20 apparently de novo CNV with significant signal gains are embedded within B73-derived regions (Figure 1). This would be expected if the apparently de novo CNV arose via SNH. Second, 12 chromosomal segments exhibited apparently de novo CNV in both RILs, resulting in 185 non-redundant segments (Table 1 and Table S2). SNH would be expected to yield a high rate of recurrent apparently de novo CNV. To further examine the degree to which the apparently de novo CNVs are shared among RILs, PCR primers were designed based on several chromosomal segments that exhibited signal loss in at least one of the RILs and used to amplify products in 300 IBM RILs. Every one of these segments was missing in multiple RILs (6%-34%) (Table S3), providing evidence that these segments exhibit frequent apparently de novo CNV origin, consistent with SNH.

The third piece of evidence that supports the role of SNH in the generation of these apparently de novo CNV is based on mapping of these sequences in B73 and Mo17 using mate pairs (Methods). Using conservative criteria 40 (M0022) and 68 (M0023) of the 197 segments exist in non-allelic positions in the B73 and Mo17 genomes (Table S4 and Table S5). In most instances (5,621/5,709) the positions of non-allelic homologs inferred by aligning the Mo17 mate-pair reads to the B73 reference genome do not correspond to expectations based on the whole genome duplication event (Krzywinski *et al.* 2009, Schnable, *et al.* 2009), indicating differential losses of genes in duplicated genomes is not the dominant mechanism underlying SNH. In contrast, only 2 (M0022) and 6 (M0023) of 197 random control segments mapped to non-allelic positions in the Mo17 genome (Figure 4 and Table S4). Hence, consistent with the SNH model, more than half of the sequences that give rise to apparently de novo CNV are located in non-allelic positions in the B73 and Mo17 genomes.

Impact of SNH-derived CNV on phenotypic traits

The SNH model predicts that meiotic segregation will act on non-allelic homologs, resulting in novel complements of sequences (losses and gains) among progeny relative to parental haplotypes (Figure 3). Those SNH-derived CNVs that involve genes would be particularly interesting because they could result in novel genic complements in progeny relative to parents.

The maize genome sequencing project defined a set of 32,540 high-quality gene annotations that is referred to as the Filtered Gene Set (FGS, ver. 4a.53). Prior to inclusion in the FGS gene models were rigorously filtered to remove gene fragments and sequences with similarity to transposons. 35 of the observed cases of SNH-derived CNVs overlap (partially or completely) with 38 of the high-quality gene models in the FGS (Table S6). RNA-Seq data from apices (GenBank accession no. SRA036595, Methods) provided evidence of expression for 24 (63%) of these genes that are affected by SNH-derived CNV (Table S6). In addition, 27 (71%) of these genes have homologs in sorghum or rice, indicating phylogenetic

conservation. In combination, these lines of evidence suggest that many of the genes affected by SNH-derived CNV are probably functional.

To test whether changes in gene complement caused by SNH-derived CNV have phenotypic consequences, we collected data on a number of phenotypic traits from the ~300 IBM RILs discussed above. Each of the 14 assayed cases of SNH-derived CNV fully or partially overlaps at least one of the high-confidence genes in the FGS. We then compared the average phenotypic trait values of RILs that did or did not experience gene loss via SNH. After controlling for multiple testing (Methods), losses of two of the 14 tested chromosomal intervals were significantly associated with phenotypic variation. Chromosomal interval M0022_seg30 is significantly associated with reduced cob diameter (adjusted p-value=0.03) and kernel row number (adjusted p-value=0.01). Similarly chromosomal interval M0022_seg15/M0023_seg22 (which includes a putative peptidyl-prolyl cis-trans isomerase gene) is associated with increased tiller number (adjusted p-value=0.01) (Table S7).

Discussion

De novo CNV has been hypothesized to arise via transposon-, recombination- and replication-mediated mechanisms (Conrad, *et al.* 2010, Hastings *et al.* 2009, Innan and Kondrashov 2010, Mills, *et al.* 2011, Springer, *et al.* 2009, Stankiewicz and Lupski 2010). The association between the distributions of gains and losses of apparently de novo CNV observed in this study and the parental origins of the surrounding chromosomal segments (Figure 1) is inconsistent with transposon-mediated mechanisms acting during the several generations required to produce the RILs. Further, the high rates of recurrence of apparently *de novo* CNV are inconsistent with recombination- and replication-driven mechanisms because these mechanisms are reported to generate losses and amplifications at much lower rates (Lupski 2007, Turner *et al.* 2008, van Ommen 2005, Yandeau-Nelson *et al.* 2006). In contrast, SNH-derived CNV is not the result of active rearrangements of DNA but is instead the result of meiotic segregation acting upon transposed gene copies and in some cases, fractionation events following the whole genome duplication. Collectively, our observations suggest that SNH results in CNV for hundreds of maize loci.

The maize genome is a product of an ancient tetraploidization event and now consists of two "subgenomes" (Schnable *et al.* 2011) having different properties, including gene expression levels. Intra-chromosomal recombination events can result in the loss of the copy of a pair of homologs from one

subgenome (Woodhouse, *et al.* 2010). Although the two subgenomes exhibit different rates of genes loss (Schnable, *et al.* 2011), this process and others such as transposon-mediated gene duplication/transpositon (Jiang, *et al.* 2004, Lai, *et al.* 2005) have generated numerous non-allelic homologs (Eichten *et al.* 2011). We have demonstrated that meiotic segregation of these non-allelic homologs generates CNV affecting hundreds of loci in the progeny of a single cross (Figure 3). The 185 detected SNH-derived CNV affect 38 high-confidence genes. Considering the stringent criteria used in this study, this frequency is likely to be an underestimate.

The SNH Model exhibits similarities to the "Reciprocal Gene Loss Model" first proposed by Lynch and Force (Lynch and Force 2000) to explain interspecific genomic incompatibility. This model proposed that the loss of different copies of duplicated genes in different populations would lead to gene loss in gametes from F1 individuals. This process has been demonstrated in crosses among three yeast species (Scannell *et al.* 2006) and between two fish species (Semon and Wolfe 2007). It has also been shown to affect single genes in several intra-specific studies, including Drosophila (Masly *et al.* 2006) and Arabidopsis (Bikard *et al.* 2009). The SNH Model differs from the Reciprocal Gene Loss Model in that it occurs intra-specifically, can generate copy number gains, and can act on non-allelic homologs generated via various mechanisms. The SNH Model would be expected to generate CNV in any species that contains non-allelic homologs and undergoes meiotic segregation. In maize we believe some of the non-allelic homologs arise via fractionation, but the mechanism outlined in Figure 3 can occur regardless of the mechanism by which the non-allelic homologs were originally generated. For example, with only minor modifications this mechanism could also generate CNVs in a species that contains non-allelic homologs generated via the transposition of single-copy genes (Vlad *et al.* 2010).

Phenotypic effects of SNH-derived CNV and PAVs

Although CNV has previously been associated with genetic disorders in humans (Stankiewicz and Lupski 2010), this report provides evidence that the segregation of CNV *via* the SNH Model can also contribute to the phenotypic variation present in crop species such as maize. This model may also shed light on transgressive segregation, i.e., the appearance of progeny from a bi-parental cross whose phenotypic values exceed those of their parents (Rieseberg *et al.* 1999).

The findings reported here have significant implications for the large-scale efforts underway to identify the genetic determinants of phenotypic variation in humans, model and agricultural species. This is because the genetic determinants of phenotypic variation arising via the SNH Model will not be detected via traditional single marker association studies or QTL analyses. Indeed, the synergistic effects from multiple unlinked genomic loci are likely to lower the power of such traditional one-dimensional analyses. Multiple-dimensional scans that consider the synergistic effects of multiple markers on

phenotypes can overcome this limitation of traditional genetic mapping approaches. Although computationally intensive, such studies are now tractable (Koesterke *et al.* 2011). Another concern is that if PAVs are not in linkage disequilibrium (LD) with nearby genetic markers such as SNPs, the power of association studies that rely on such markers will be reduced. It will be interesting to determine whether the direct genotyping of PAVs via CGH-based genotyping (Fu, *et al.* 2010) or genotyping-by-sequencing approaches (Andolfatto *et al.* 2011, Elshire *et al.* 2011, Fogu *et al.* 2007, Huang *et al.* 2009, Lai, *et al.* 2010) will uncover at least a fraction of the "missing heritability" observed in genome-wide association studies (Kump *et al.* 2010).

Material & Methods

Genetic stocks

Two maize inbred lines, B73 and Mo17, and two recombinant inbred lines (RILs), M0022 and M0023, are extracted from the Intermated B73 x Mo17 (IBM) Syn4 population (Lee, *et al.* 2002). The RILs used in this study were from the F7-9 generation.

Identification of putative de novo CNV probes

The array CGH experiments (GEO: GSE16938), data processing and statistical analyses were performed as described previously (Fu, *et al.* 2010, Springer, *et al.* 2009). Contrasts were performed between B73 vs. Mo17, B73 vs. RIL and Mo17 vs. RIL. A p-value was determined for each probe from each contrast. To account for multiple testing, p-values were converted to q-values (Benjamini and Hochberg 1995). Probes with significantly higher or lower signals in the RILs when compared to both B73 and Mo17 were termed putative *de novo* CNV probes. Signal loss probes were called using the criteria of q-value (RIL vs. B73)<0.0001, q-value (RIL vs. Mo17)<0.0001, log2(signal ratios of RIL/B73)<0 and log2(signal ratios of RIL/Mo17)<0; signal gain probes were called using the criteria of q-value (RIL vs. B73)<0.001, q-value (RIL vs. Mo17)<0.001, log2(signal ratios of RIL/B73)>0 and log2(signal ratios of RIL/Mo17)>0. Different q-value cutoffs were used to identify signal loss probes and signal gain probes because aCGH technology has a greater power to detect copy number losses than copy number gains (Altshuler *et al.* 2010).

Determination of appropriate parental control for each probe in aCGH analyses of RILs

For each aCGH probe, the log2 ratios of the hybridization signals of the RIL vs. B73 (log2(RIL/B)) and separately vs. Mo17 (log2(RIL/M)) were calculated. A log2 ratio that is greater (or smaller) than 0

indicates that a given probe yields a stronger (weaker) signal in the RIL than in a particular parental inbred. For each aCGH probe, the smaller of absolute value of log2(RIL/B) and the absolute value of log2(RIL/M) was used to identify the presumptive parental origin (B73 or Mo17) in the RIL of the chromosomal segment from which each probe was designed. For each probe this parental hybridization value (log2(RIL/B|M) was used for the calculations plotted in Figure 1.

Segmentation of putative *de novo* CNV probes to identify putative de novo CNV segments

The putative *de novo* CNV probes were converted to 1 (copy gain compared to B73) or -1 (copy loss compared to B73). All other probes were assigned a value of 0. The converted binary data were subjected to segmentation via DNAcopy (Olshen *et al.* 2004) using the parameters: alpha=0.01, nperm=10000, p.method="perm", eta=0.01, min.width=3. Putative *de novo* CNV segments were required to contain at least three putative *de novo* CNV probes and a median absolute deviation (MAD) equal to 0.

RIL segmentation to distinguish the origin of regions from either B73 or Mo17

Probes that distinguished B73 and Mo17 were identified using the criteria: q-value (Mo17 vs. B73)<0.0001 & log2(Mo17/B73) < (-1). These probes were treated as genetic markers to genotype the RILs, most of which were grouped into Mo17-type (q-value (RIL vs. B73)<0.001 & q-value (RIL vs. Mo17)>0.1) or B73-type (q-value (RIL vs. B73)>0.1 & q-value (RIL vs Mo17)<0.001).

To avoid the partition of a chromosomal region exhibiting the same origin into multiple segments, the array CGH genotyping results were converted to binary data (B73-type=1; Mo17-type=0) and merged to perform segmentation as described above. Segments smaller than 200 kb and having a mean segment value between 0.1 and 0.9 were removed from further analysis. By so doing, we excluded segments that were ambiguously assigned as being B73-type or Mo17-type and small segments that might represent misassemblies in the reference genome.

Exome-Seq

Gene annotation information was downloaded from www.maizesequence.org (http://ftp.maizesequence.org/release-4a.53/filtered-set/ZmB73_4a.53_FGS.gff.gz). This annotation set was defined as the entire set of evidence-based genes (predicted by Gramene GeneBuilder), complemented by a set of Fgenesh models. Pseudogenes, TE-encoded genes, and low-confidence models were filtered out to produce the final annotation set. Coding sequence coordinates were extracted from this annotation (251,067 regions; 55.5Mbp) and consolidated into non-overlapping regions (152,529 regions; 37.9Mbp). These regions were padded to a minimum region size of 100bp, and again

consolidated into non-overlapping regions (151,929 regions; 39.7Mbp). Final coordinates were offset by 35bp to account for capture probe length overhang at the end of each region. Variable length capture probes (50- to 100-mers) were selected by tiling through each region at an average spacing of 48bp (measured from 5' start to 5' start). Repeat-masking was done by generating a histogram of all 15-mers in the maize genome and removing probes with an average 15-mer frequency greater than 100. Probe uniqueness was assessed using SSAHA (http://www.sanger.ac.uk/resources/software/ssaha/), using a minimum match size of 26. No more than 3 matches were allowed for capture probes. The final design covers 131,469 of the 151,929 original regions, and has a total capture space of 55.4Mbp. The Sequence Capture Developer 2.1M feature array design 100224_ZmB73_public_exome_cap_HX3 is available for purchase from Roche NimbleGen (http://www.nimblegen.com/).

Analysis of Exome-Seq data

Novoalign 2.05.31 (http://www.novocraft.com/) was used to map all reads (40bp) to reference genomes, including the B73 reference genome (B73ref_v1), the maize mitochondrial genome (Genbank acc#: AY506529.1) and the maize chloroplast genome (Genbank acc#: X86563.2). Reads that were uniquely mapped with≤2 mismatches (insertions and deletions was counted as mismatches) were used for further analysis. Read counts of each putative *de novo* CNV segment were adjusted by the addition of 1 to avoid zero values and were then used to calculate log2(RIL/B73) and log2(RIL/Mo17).

RNA-Seq

RNA was extracted from a pool of 3-6 apexes of 14-day old seedlings of the inbreds B73 and Mo17 using the Qiagen RNeasy Plant Mini Kit (Cat. # 74903, http://www.qiagen.com/default.aspx). RNA-Seq was conducted using an Illumina GAIIx instrument at the Iowa State University DNA facility following an Illumina protocol (mRNA-seq Sample Preparation Guide, http://www.illumina.com/).

Genotyping of a subset of the apparently *de novo* CNV segments on the RIL population

Primers were designed on a subset of the apparently *de novo* CNV segments that fully or partially overlap with transcription- or annotation-supported genes to directly genotype >300 intermated B73xMo17 recombinant inbred lines. A primer pair was designed for each selected segment using the B73 reference sequence. Primers were aligned to JGI 454 Mo17 reads to ensure that primers work for Mo17. This PCR program consisted of 94°C for 10 min; 35 cycles of 94°C for 30 sec, 60°C for 45 sec, 72°C for 1.5 min; and a final extension at 72°C for 10 min in a 20-□l volume. Three existing genetic markers of the intermated B73xMo17 genetic map (Liu, *et al.* 2009), IDP525, IDP7957 and IDP7324, were used as the PCR control primers. The IDP525 marker has been used to genotype the full set of RILs previously (Liu, *et al.* 2009). The RILs with poor or no PCR amplification of any IDP markers or inconsistent IDP525

PCR results between the previous scores and the re-genotyping scores were not used for further analysis.

Phenotyping RILs

The following traits were collected from 291 IBM RILs: seedling dry weight, average kernel weight, cob diameter, cob length, cob weight, kernel count, kernel row number, total kernel weight, tiller number, ferulic acid (FA), *p*-coumaric acid (PCA), brace node number (BN), node number above primary ear (NA), node number below primary ear (NB) and total node number (NT). Four replicates of seedling dry weight trait data and three replications of other traits data were collected. Least square means for each genotype were estimated using the SAS code for the traits except for tiller number. Mean of tiller number for each genotype was calculated among replicates.

Phenotypic associations of SNH-derived CNVs

T-tests were used to test the null hypothesis that no association exists between segmental copy loss and each of the phenotypic traits. For each tested segment, RILs were divided into two groups based on the results of PCR-based genotyping, i.e., RILs with and without the expected PCR bands. The t-test assuming equal variation between two groups was conducted for each trait, generating a p-value. A permutation test was implemented to account for multiple testing for each segmental copy loss. In each permutation of a given segment, the genotyping scores of the segment were randomly shuffled among the RILs. A similar t-test was performed on the shuffled genotyping data for each trait and a p-value was obtained for each trait. The smallest p-value was then determined among the multiple p-values from the multiple traits' tests. This procedure was repeated 1,000 times. 1,000 p-values were obtained for each of the tested segments. The original p-values of the multiple traits computed from the observed data of a given segment were compared to the 1000 p-values from the permutation test of this segment to generate adjusted p-values.

Physical mapping of Mo17 sequences involved in apparently *de novo* CNV

To determine the chromosomal location of paired reads, the mate-pair cluster mapping method was used. First, the individual reads from 5-kb Mo17 mate-pairs (1 lane from an Illumina/Solexa GA-IIx instrument) were separately mapped to the B73 reference genome. A mate pair was used for further analyses only if one and only one member of that mate pair uniquely mapped to an apparently *de novo* CNV segment (or within the 500 bp beyond either end of the segment because the actual endpoints of loss or duplication are not known with certainty) in the B73 reference genome. For each mate-pair of this type, if the other member mapped within 1 Mb of the same apparently *de novo* CNV segment, the pair was categorized as having similar locations in the B73 and Mo17 genomes ("locally mapped"). If the mapped location of the other member of a mate-pair was >1 Mb away or on a different chromosome, the mate-pair

was categorized as having different locations in the B73 and Mo17 genomes ("distally mapped"). The identification of multiple, independent distally-mapped mate-pairs clustered on a chromosomal region was considered evidence that the corresponding chromosomal segments in B73 and Mo17 are non-allelic. A "cluster" of mate pair reads was defined as consisting of ≥2 independent non-stacked reads that mapped within <100 kb of each other and that covered >80 bp of the CNV segment. Two sets of segments (one from M00022 and one from M0023) were randomly selected as the controls. For each RIL, the number of randomly selected segments from each chromosome was equal to the number of identified apparently *de novo* CNV segments on that chromosome in the corresponding RIL.

Acknowledgements

We thank Stephen Moldovan, Ho Man Tang and Marianne Smith for technical assistance and Tracy Millard and Dr. Thomas J. Albert for sequencing and project support. We thank Roche NimbleGen for providing support for this study through the donation of reagents, and team-member time. The co-author Todd Richmond, Daniel J. Gerhardt and Jeffrey A. Jeddeloh recognize a competing interest in this publication as employees of Roche NimbleGen, Inc. This research was also supported by a grant from the National Science Foundation to P.S.S. (IOS-1027527) and data generated as part of NSF grant IOS-0820610 (Mike Scanlon, PI).

References

Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., Bonnen, P.E., de Bakker, P.I.C.-L., Deloukas, P.C.-L., Gabriel, S.B., Gwilliam, R., Hunt, S., Inouye, M.C.-L., Jia, X., Palotie, A., Parkin, M.C.-L., Whittaker, P., Yu, F.L., Chang, K., Hawes, A., Lewis, L.R., Ren, Y., Wheeler, D., Marie Muzny, D., Barnes, C., Darvishi, K., Hurles, M.C.-L., Korn, J.M., Kristiansson, K., Lee, C., McCarroll, S.A.C.-L., Nemesh, J., Keinan, A.L., Montgomery, S.B., Pollack, S., Price, A.L., Soranzo, N., Gonzaga-Jauregui, C., Keinan, A., Anttila, V., Brodeur, W., Daly, M.J., Leslie, S., McVean, G., Moutsianas, L., Nguyen, H., Schaffner, S.F.L., Zhang, Q., Ghori, M.J., McGinnis, R.C.-L., McLaren, W., Price, A.L.C.-L., Schaffner, S.F.C.-L., Takeuchi, F., Grossman, S.R., Shlyakhter, I., Hostetter, E.B., Sabeti, P.C.L., Adebamowo, C.A., Foster, M.W., Gordon, D.R., Licinio, J., Cristina Manca, M., Marshall, P.A., Matsuda, I., Ngare, D., Ota Wang, V., Reddy, D., Rotimi, C.N., Royal, C.D., Sharp, R.R., Zeng, C., Brooks, L.D. and McEwen, J.E. (2010) Integrating common and rare genetic variation in diverse human populations. *Nature*, **467**, 52-58.

Andolfatto, P., Davison, D., Erezyilmaz, D., Hu, T.T., Mast, J., Sunayama-Morita, T. and Stern, D.L. (2011) Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Res*, **21**, 610-617.

Belo, A., Beatty, M.K., Hondred, D., Fengler, K.A., Li, B. and Rafalski, A. (2010) Allelic genome structural variations in maize detected by array comparative genome hybridization. *Theor Appl Genet*, **120**, 355-367.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statistical Society, Series B*, **57**, 289–300.

Bennetzen, J.L. (2000) Transposable element contributions to plant gene and genome evolution. *Plant Mol Biol*, **42**, 251-269.

Bikard, D., Patel, D., Le Mette, C., Giorgi, V., Camilleri, C., Bennett, M.J. and Loudet, O. (2009) Divergent evolution of duplicate genes leads to genetic incompatibilities within A. thaliana. *Science*, **323**, 623-626.

Brunetti-Pierri, N., Berg, J.S., Scaglia, F., Belmont, J., Bacino, C.A., Sahoo, T., Lalani, S.R., Graham, B., Lee, B., Shinawi, M., Shen, J., Kang, S.H., Pursley, A., Lotze, T., Kennedy, G., Lansky-Shafer, S., Weaver, C., Roeder, E.R., Grebe, T.A., Arnold, G.L., Hutchison, T., Reimschisel, T., Amato, S., Geragthy, M.T., Innis, J.W., Obersztyn, E., Nowakowska, B., Rosengren, S.S., Bader, P.I., Grange, D.K., Naqvi, S., Garnica, A.D., Bernes, S.M., Fong, C.T., Summers, A., Walters, W.D., Lupski, J.R., Stankiewicz, P., Cheung, S.W. and Patel, A. (2008) Recurrent reciprocal 1q21.1 deletions and duplications associated with microcephaly or macrocephaly and developmental and behavioral abnormalities. *Nat Genet*, **40**, 1466-1471.

Chen, W.K., Swartz, J.D., Rush, L.J. and Alvarez, C.E. (2009) Mapping DNA structural variation in dogs. *Genome Res*, **19**, 500-509.

Conrad, D.F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T.D., Barnes, C., Campbell, P., Fitzgerald, T., Hu, M., Ihm, C.H., Kristiansson, K., Macarthur, D.G., Macdonald, J.R., Onyiah, I., Pang, A.W., Robson, S., Stirrups, K., Valsesia, A., Walter, K., Wei, J., Tyler-Smith, C., Carter, N.P., Lee, C., Scherer, S.W. and Hurles, M.E. (2010) Origins and functional impact of copy number variation in the human genome. *Nature*, **464**, 704-712.

Daines, B., Wang, H., Li, Y., Han, Y., Gibbs, R. and Chen, R. (2009) High-throughput multiplex sequencing to discover copy number variants in Drosophila. *Genetics*, **182**, 935-941.

DeBolt, S. (2010) Copy number variation shapes genome diversity in Arabidopsis over immediate family generational scales. *Genome Biol Evol*, **2**, 441-453.

Eichten, S.R., Foerster, J.M., de Leon, N., Kai, Y., Yeh, C.T., Liu, S., Jeddeloh, J.A., Schnable, P.S., Kaeppler, S.M. and Springer, N.M. (2011) B73-Mo17 near-isogenic lines demonstrate dispersed structural variation in maize. *Plant Physiol*, **156**, 1679-1690.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S. and Mitchell, S.E. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*, **6**, e19379.

Fadista, J., Nygaard, M., Holm, L.E., Thomsen, B. and Bendixen, C. (2008) A snapshot of CNVs in the pig genome. *PLoS One*, **3**, e3916.

Fadista, J., Thomsen, B., Holm, L.E. and Bendixen, C. (2010) Copy number variation in the bovine genome. *BMC Genomics*, **11**, 284.

Fernandez, B.A., Roberts, W., Chung, B., Weksberg, R., Meyn, S., Szatmari, P., Joseph-George, A.M., Mackay, S., Whitten, K., Noble, B., Vardy, C., Crosbie, V., Luscombe, S., Tucker, E., Turner, L., Marshall, C.R. and Scherer, S.W. (2010) Phenotypic spectrum associated with de novo and inherited deletions and duplications at 16p11.2 in individuals ascertained for diagnosis of autism spectrum disorder. *J Med Genet*, **47**, 195-203.

Fogu, G., Bandiera, P., Cambosu, F., Carta, A.R., Pilo, L., Serra, G., Soro, G., Tondi, M., Tusacciu, G. and Montella, A. (2007) Pure partial trisomy of 6p12.1-p22.1 secondary to a familial 12/6 insertion in two malformed babies. *Eur J Med Genet*, **50**, 103-111.

Fu, Y., Springer, N.M., Ying, K., Yeh, C.T., Iniguez, A.L., Richmond, T., Wu, W., Barbazuk, B., Nettleton, D., Jeddeloh, J. and Schnable, P.S. (2010) High-resolution genotyping via whole genome hybridizations to microarrays containing long oligonucleotide probes. *PLoS One*, **5**, e14178.

Hastings, P.J., Ira, G. and Lupski, J.R. (2009) A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genet*, **5**, e1000327.

Haun, W.J., Hyten, D.L., Xu, W.W., Gerhardt, D.J., Albert, T.J., Richmond, T., Jeddeloh, J.A., Jia, G., Springer, N.M., Vance, C.P. and Stupar, R.M. (2010) The composition and origins of genomic

variation among individuals of the soybean reference cultivar Williams 82. *Plant Physiol*, **155**, 645-655.

Huang, X., Feng, Q., Qian, Q., Zhao, Q., Wang, L., Wang, A., Guan, J., Fan, D., Weng, Q., Huang, T., Dong, G., Sang, T. and Han, B. (2009) High-throughput genotyping by whole-genome resequencing. *Genome Res*, **19**, 1068-1076.

Hurwitz, B.L., Kudrna, D., Yu, Y., Sebastian, A., Zuccolo, A., Jackson, S.A., Ware, D., Wing, R.A. and Stein, L. (2010) Rice structural variation: a comparative analysis of structural variation between rice and three of its closest relatives in the genus Oryza. *Plant J*, **63**, 990-1003.

Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W. and Lee, C. (2004) Detection of large-scale variation in the human genome. *Nat Genet*, **36**, 949-951.

Innan, H. and Kondrashov, F. (2010) The evolution of gene duplications: classifying and distinguishing between models. *Nat Rev Genet*, **11**, 97-108.

Jiang, N., Bao, Z., Zhang, X., Eddy, S.R. and Wessler, S.R. (2004) Pack-MULE transposable elements mediate gene evolution in plants. *Nature*, **431**, 569-573.

Koesterke, L., Stanzione, D., Vaughn, M., Welch, S.M., Kusnierczyk, W., Wang, J., Yeh, C., Nettleton, D. and Schnable, P.S. (2011) An efficient and scalable implementation of SNP-pair interaction testing for genetic association studies. In *IEEE International Workshop on High Performance Computational Biology*. Anchorage, Alaska.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J. and Marra, M.A. (2009) Circos: an information aesthetic for comparative genomics. *Genome Res*, **19**, 1639-1645.

Kump, K.L., Bradbury, P.J., Wisser, R.J., Buckler, E.S., Belcher, A.R., Oropeza-Rosas, M.A., Zwonitzer, J.C., Kresovich, S., McMullen, M.D., Ware, D., Balint-Kurti, P.J. and Holland, J.B. (2010) Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. *Nat Genet*, **43**, 163-168.

Lai, J., Li, R., Xu, X., Jin, W., Xu, M., Zhao, H., Xiang, Z., Song, W., Ying, K., Zhang, M., Jiao, Y., Ni, P., Zhang, J., Li, D., Guo, X., Ye, K., Jian, M., Wang, B., Zheng, H., Liang, H., Zhang, X., Wang, S., Chen, S., Li, J., Fu, Y., Springer, N.M., Yang, H., Wang, J., Dai, J. and Schnable, P.S. (2010) Genome-wide patterns of genetic variation among elite maize inbred lines. *Nat Genet*, **42**, 1027-1030.

Lai, J., Li, Y., Messing, J. and Dooner, H.K. (2005) Gene movement by Helitron transposons contributes to the haplotype variability of maize. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 9068-9073.

Langham, R.J., Walsh, J., Dunn, M., Ko, C., Goff, S.A. and Freeling, M. (2004) Genomic duplication, fractionation and the origin of regulatory novelty. *Genetics*, **166**, 935-945.

Lee, M., Sharopova, N., Beavis, W.D., Grant, D., Katt, M., Blair, D. and Hallauer, A. (2002) Expanding the genetic map of maize with the intermated B73 x Mo17 (IBM) population. *Plant Mol Biol*, **48**,

453-461.

Liu, S., Yeh, C.T., Ji, T., Ying, K., Wu, H., Tang, H.M., Fu, Y., Nettleton, D. and Schnable, P.S. (2009) Mu transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome. *PLoS Genet*, **5**, e1000733.

Lu, P., Han, X., Qi, J., Yang, J., Wijeratne, A.J., Li, T. and Ma, H. (2012) Analysis of Arabidopsis genome-wide variations before and after meiosis and meiotic recombination by resequencing Landsberg erecta and all four products of a single meiosis. *Genome Res*, **22**, 508-518.

Lupski, J.R. (2007) Genomic rearrangements and sporadic disease. *Nat Genet*, **39**, S43-47.

Lynch, M. and Force, A. (2000) The origin of interspecific genomic incompatibility via gene duplication. *Am Nat*, **156**, 590-605.

Masly, J.P., Jones, C.D., Noor, M.A., Locke, J. and Orr, H.A. (2006) Gene transposition as a cause of hybrid sterility in Drosophila. *Science*, **313**, 1448-1450.

Mills, R.E., Walter, K., Stewart, C., Handsaker, R.E., Chen, K., Alkan, C., Abyzov, A., Yoon, S.C., Ye, K., Cheetham, R.K., Chinwalla, A., Conrad, D.F., Fu, Y., Grubert, F., Hajirasouliha, I., Hormozdiari, F., Iakoucheva, L.M., Iqbal, Z., Kang, S., Kidd, J.M., Konkel, M.K., Korn, J., Khurana, E., Kural, D., Lam, H.Y., Leng, J., Li, R., Li, Y., Lin, C.Y., Luo, R., Mu, X.J., Nemesh, J., Peckham, H.E., Rausch, T., Scally, A., Shi, X., Stromberg, M.P., Stutz, A.M., Urban, A.E., Walker, J.A., Wu, J., Zhang, Y., Zhang, Z.D., Batzer, M.A., Ding, L., Marth, G.T., McVean, G., Sebat, J., Snyder, M., Wang, J., Eichler, E.E., Gerstein, M.B., Hurles, M.E., Lee, C., McCarroll, S.A. and Korbel, J.O. (2011) Mapping copy number variation by population-scale genome sequencing. *Nature*, **470**, 59-65.

Neill, N.J., Ballif, B.C., Lamb, A.N., Parikh, S., Ravnan, J.B., Schultz, R.A., Torchia, B.S., Rosenfeld, J.A. and Shaffer, L.G. (2011) Recurrence, submicroscopic complexity, and potential clinical relevance of copy gains detected by array CGH that are shown to be unbalanced insertions by FISH. *Genome Res*, **21**, 535-544.

Olshen, A.B., Venkatraman, E.S., Lucito, R. and Wigler, M. (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557-572.

Rieseberg, L.H., Archer, M.A. and Wayne, R.K. (1999) Transgressive segregation, adaptation and speciation. *Heredity*, **83**, 363-372.

Scannell, D.R., Byrne, K.P., Gordon, J.L., Wong, S. and Wolfe, K.H. (2006) Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature*, **440**, 341-345.

Schnable, J.C., Springer, N.M. and Freeling, M. (2011) Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 4069-4074.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A., Minx, P., Reily, A.D., Courtney, L., Kruchowski, S.S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., Rock, S.M., Belter, E., Du, F., Kim, K., Abbott, R.M.,

Cotton, M., Levy, A., Marchetto, P., Ochoa, K., Jackson, S.M., Gillam, B., Chen, W., Yan, L., Higginbotham, J., Cardenas, M., Waligorski, J., Applebaum, E., Phelps, L., Falcone, J., Kanchi, K., Thane, T., Scimone, A., Thane, N., Henke, J., Wang, T., Ruppert, J., Shah, N., Rotter, K., Hodges, J., Ingenthron, E., Cordes, M., Kohlberg, S., Sgro, J., Delgado, B., Mead, K., Chinwalla, A., Leonard, S., Crouse, K., Collura, K., Kudrna, D., Currie, J., He, R., Angelova, A., Rajasekar, S., Mueller, T., Lomeli, R., Scara, G., Ko, A., Delaney, K., Wissotski, M., Lopez, G., Campos, D., Braidotti, M., Ashley, E., Golser, W., Kim, H., Lee, S., Lin, J., Dujmic, Z., Kim, W., Talag, J., Zuccolo, A., Fan, C., Sebastian, A., Kramer, M., Spiegel, L., Nascimento, L., Zutavern, T., Miller, B., Ambroise, C., Muller, S., Spooner, W., Narechania, A., Ren, L., Wei, S., Kumari, S., Faga, B., Levy, M.J., McMahan, L., Van Buren, P., Vaughn, M.W., Ying, K., Yeh, C.T., Emrich, S.J., Jia, Y., Kalyanaraman, A., Hsia, A.P., Barbazuk, W.B., Baucom, R.S., Brutnell, T.P., Carpita, N.C., Chaparro, C., Chia, J.M., Deragon, J.M., Estill, J.C., Fu, Y., Jeddeloh, J.A., Han, Y., Lee, H., Li, P., Lisch, D.R., Liu, S., Liu, Z., Nagel, D.H., McCann, M.C., SanMiguel, P., Myers, A.M., Nettleton, D., Nguyen, J., Penning, B.W., Ponnala, L., Schneider, K.L., Schwartz, D.C., Sharma, A., Soderlund, C., Springer, N.M., Sun, Q., Wang, H., Waterman, M., Westerman, R., Wolfgruber, T.K., Yang, L., Yu, Y., Zhang, L., Zhou, S., Zhu, Q., Bennetzen, J.L., Dawe, R.K., Jiang, J., Jiang, N., Presting, G.G., Wessler, S.R., Aluru, S., Martienssen, R.A., Clifton, S.W., McCombie, W.R., Wing, R.A. and Wilson, R.K. (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science*, **326**, 1112-1115.

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Maner, S., Massa, H., Walker, M., Chi, M., Navin, N., Lucito, R., Healy, J., Hicks, J., Ye, K., Reiner, A., Gilliam, T.C., Trask, B., Patterson, N., Zetterberg, A. and Wigler, M. (2004) Large-scale copy number polymorphism in the human genome. *Science*, **305**, 525-528.

Semon, M. and Wolfe, K.H. (2007) Reciprocal gene loss between Tetraodon and zebrafish after whole genome duplication in their ancestor. *Trends Genet*, **23**, 108-112.

She, X., Cheng, Z., Zollner, S., Church, D.M. and Eichler, E.E. (2008) Mouse segmental duplication and copy number variation. *Nat Genet*, **40**, 909-914.

Shinawi, M., Liu, P., Kang, S.H., Shen, J., Belmont, J.W., Scott, D.A., Probst, F.J., Craigen, W.J., Graham, B.H., Pursley, A., Clark, G., Lee, J., Proud, M., Stocco, A., Rodriguez, D.L., Kozel, B.A., Sparagana, S., Roeder, E.R., McGrew, S.G., Kurczynski, T.W., Allison, L.J., Amato, S., Savage, S., Patel, A., Stankiewicz, P., Beaudet, A.L., Cheung, S.W. and Lupski, J.R. (2010) Recurrent reciprocal 16p11.2 rearrangements associated with global developmental delay, behavioural problems, dysmorphism, epilepsy, and abnormal head size. *J Med Genet*, **47**, 332-341.

Springer, N.M., Ying, K., Fu, Y., Ji, T., Yeh, C.T., Jia, Y., Wu, W., Richmond, T., Kitzman, J., Rosenbaum, H., Iniguez, A.L., Barbazuk, W.B., Jeddeloh, J.A., Nettleton, D. and Schnable, P.S. (2009) Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV)

in genome content. *PLoS Genet*, **5**, e1000734.

Stankiewicz, P. and Lupski, J.R. (2010) Structural variation in the human genome and its role in disease. *Annu Rev Med*, **61**, 437-455.

Swanson-Wagner, R.A., Eichten, S.R., Kumari, S., Tiffin, P., Stein, J.C., Ware, D. and Springer, N.M. (2010) Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome Res*, 1689-1699.

Turner, D.J., Miretti, M., Rajan, D., Fiegler, H., Carter, N.P., Blayney, M.L., Beck, S. and Hurles, M.E. (2008) Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. *Nat Genet*, **40**, 90-95.

van Ommen, G.J. (2005) Frequency of new copy number variation in humans. *Nat Genet*, **37**, 333-334.

Vlad, D., Rappaport, F., Simon, M. and Loudet, O. (2010) Gene transposition causing natural variation for growth in Arabidopsis thaliana. *PLoS Genet*, **6**, e1000945.

Woodhouse, M.R., Schnable, J.C., Pedersen, B.S., Lyons, E., Lisch, D., Subramaniam, S. and Freeling, M. (2010) Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homologs. *PLoS biology*, **8**, e1000409.

Yandeau-Nelson, M.D., Xia, Y., Li, J., Neuffer, M.G. and Schnable, P.S. (2006) Unequal sister chromatid and homolog recombination at a tandem duplication of the A1 locus in maize. *Genetics*, **173**, 2211-2226.

Zhang, J. and Peterson, T. (2005) A segmental deletion series generated by sister-chromatid transposition of Ac transposable elements in maize. *Genetics*, **171**, 333-344.

Tables

Table 1. Numbers of putative *de novo* CNV segments (genes in putative *de novo* CNV segments) derived from M0022 and M0023

	M0023 Gain	**M0023 Loss**	**No gain/loss in M0023**	**Total**
M0022 Gain	2 (3)	0 (0)	2 (0)	4 (3)
M0022 Loss	3 (3)	7 (1)	53 (10)	63 (14)
No gain/loss in M0022	11 (2)	107 (19)	0 (0)	118 (21)
Total	16 (8)	114 (20)	55 (10)	185 (38)

Figures

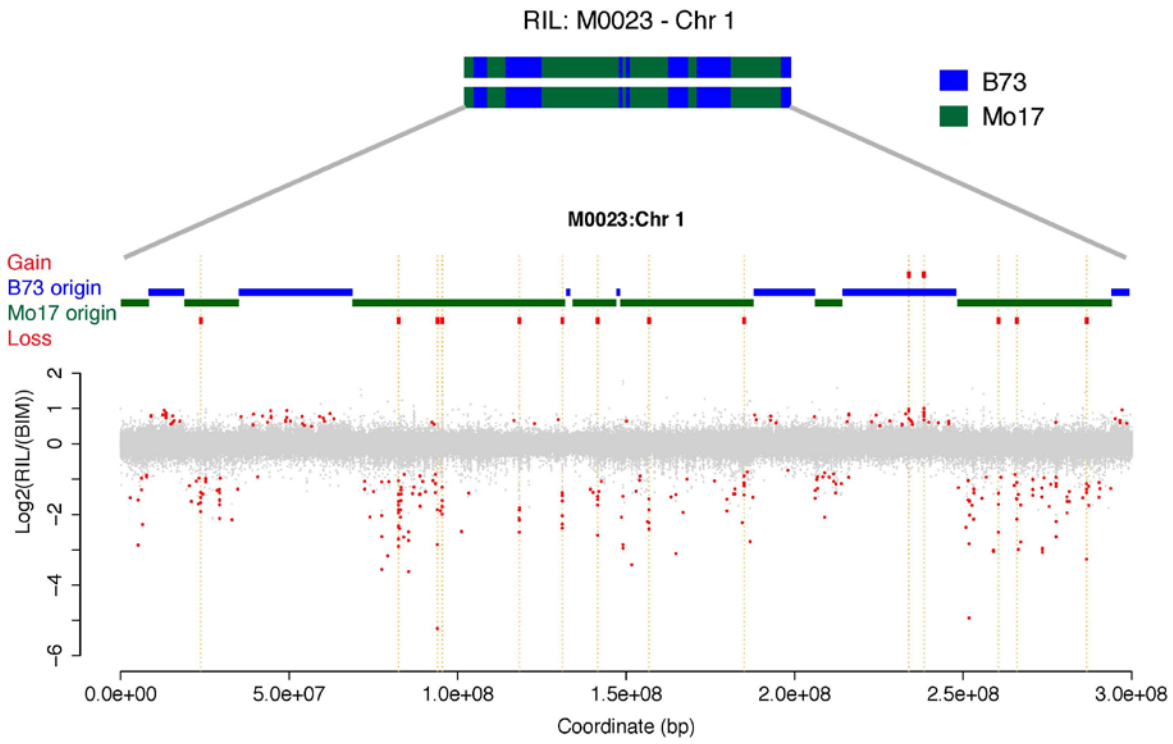


Figure 1

Figure 1. Distribution of putative de novo CNV on chromosome 1. The physical chromosomal position of each aCGH probe is plotted versus the log2 ratio of its hybridization signals in RIL M0023 and the appropriate parent (i.e., B73 or Mo17, (log2(RIL/B|M) see Methods). Probes that did or did not exhibit statistically significant signal losses or gains relative to both parents are highlighted in red and grey, respectively. Above the X-Y plot, the chromosomal regions from RIL M0023 that were derived from B73 and Mo17 based on genotyping experiments are color-coded in blue and green, respectively. These marker-based assignments of parental origin of chromosomal regions were inferred via segmentation of all aCGH probes that could be reliably classified as having B73-like or Mo17-like signals. Subsequently, a second segmentation was conducted to identify putative de novo CNV segments that are indicated in red.

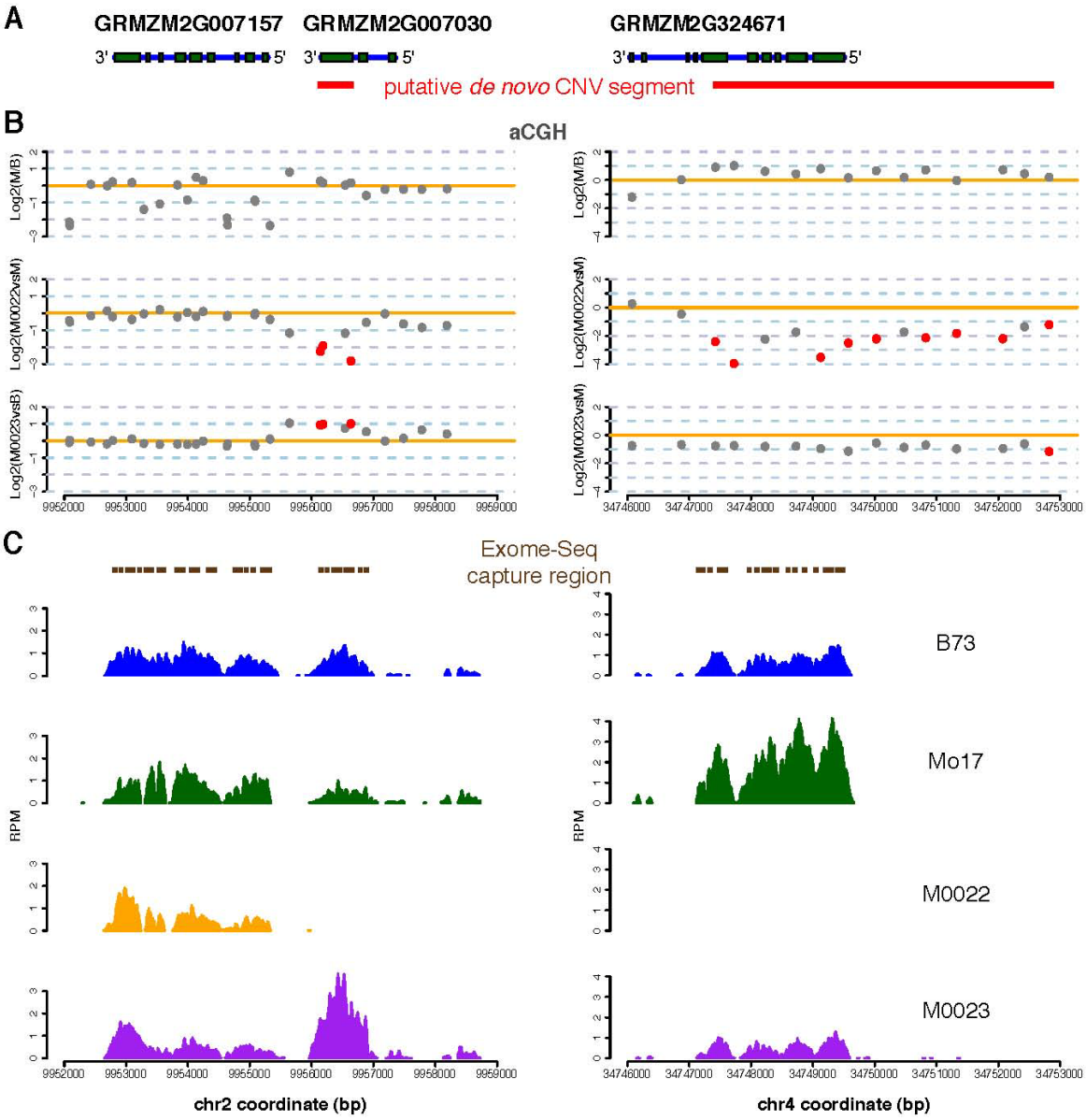


Figure 2

Figure 2. Analysis of two apparently *de novo* CNV segments via aCGH and exome-Seq. (A) Coding regions of three genes from two genomic regions are illustrated by green boxes. (B) The log2 of the ratios of normalized probe signals between different genotypes from aCGH are plotted versus probe's physical positions. Probes that did or did not exhibit statistically significant signal losses or gains relative to both parents are highlighted in red and grey, respectively. Probes in red represent putative *de novo* CNV probes. Comparisons were conducted between the RILs and the parent that contributed the relevant chromosomal segment. (C) Read counts from the exome capture experiment at each nucleotide position were normalized to reads per million reads (RPM) for each genotype.

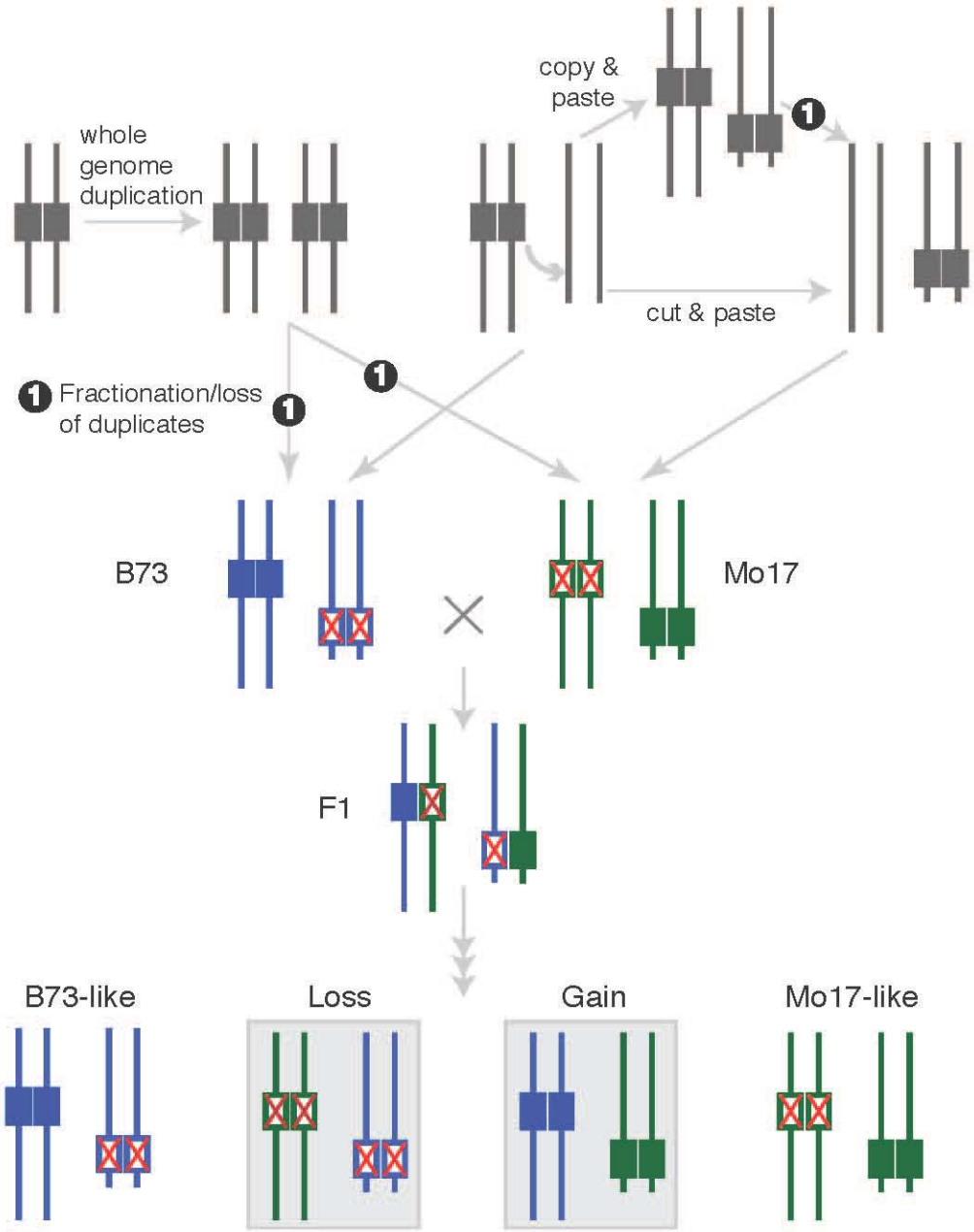


Figure 3

Figure 3. The SNH Model for the origin of recurrent apparently de novo CNV. Identical or nearly identical DNA sequences are located on different chromosomes in the B73 and Mo17 genomes (non-allelic homologs). B73- and Mo17-derived chromosomes are indicated by blue and green, respectively. Filled boxes and open boxes containing a red "X" designate the presence and absence of a non-allelic homolog, respectively. This model is consistent with our finding that most recurrent apparently de novo CNVs exhibit losses in ~25% (2 loci) or ~12.5% (3 loci) of the RILs. It is also consistent with our finding that all of the 165 losses are embedded in Mo17-derived segments and that almost all (19/20) of the copy number gains are embedded in B73-derived segments (Figure 1). This model predicts that copy number losses and gain should occur at equal frequencies. We hypothesize that losses exceed gains due to ascertainment bias (i.e., copy number losses are more easily detected than copy number gains).

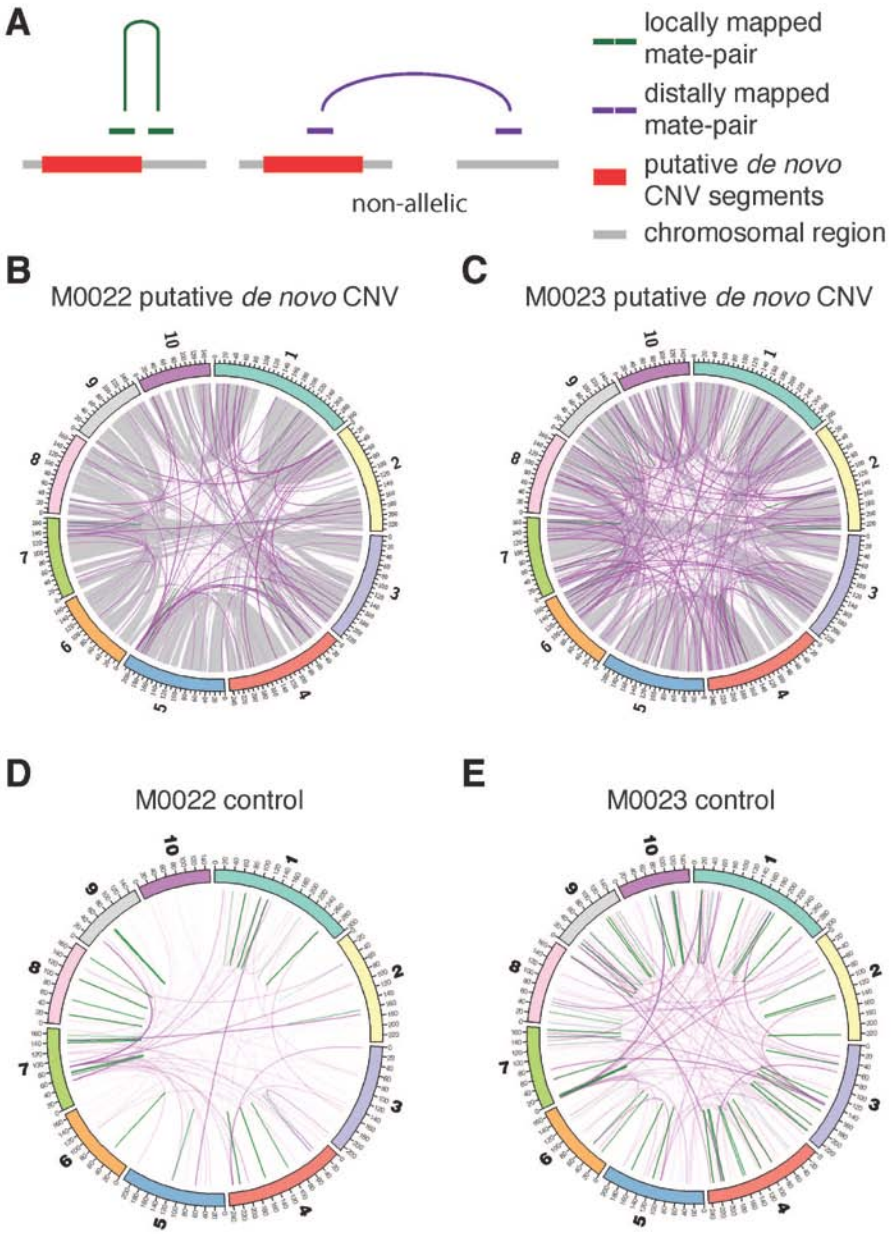


Figure 4

Figure 4. Determination of the allelic status in the B73 and Mo17 genomes of sequences that give rise to apparently *de novo* CNV.

(A) Mo17 mate-pair reads with 5-kb of insertion were mapped separately to the B73 reference genome. Pairs with at least one read uniquely mapped to the apparently de novo CNV segments (red horizontal bars) and the other read were uniquely mapped to the reference genome were categorized to two types, locally mapped mate-pair (green horizontal bars) and distally mapped mate-pair (purple horizontal bars). Locally mapped mate-pair linked by a green line represents mate-pair reads that were mapped within 1Mb, while distally mapped mate-pair linked by a purple line represents mate-pair reads that were mapped at >1Mb distance or on different chromosomes. These rules were applied to draw

figures in the panels B, C, D, and E. (B-E) Mate-pair reads associated with the apparently de novo CNV segments and control segments from the indicated genotypes are color-coded as shown in panel A. Each line connects mapped locations of two Mo17 reads of each mate-pair at least one of which was uniquely mapped to the apparently de novo CNV segments of M0022 (B) and the apparently de novo CNV segments of M0023 (C), randomly selected segments simulating M0022 apparently de novo CNV segments (D) and randomly selected segments simulating M0023 apparently de novo CNV segments (E). The number of read pairs clustered was intensity-coded as shown in panel (A). The transparency factor for each line is 0.05. In panels B and C, the homoeologous sites of duplicated blocks derived from the ancestral whole genome duplication in the B73 reference genome are shown in grey.

CHAPTER 5. GENERAL CONCLUSIONS AND PROSPECTS OF POPULATION LEVEL ANALYSES OF STRUCTURAL VARIATION IN *ZEA*

Maize (*Zea mays* L. *ssp. mays*) is both an important crop and a model for genetic and genomic studies. The maize genome is highly repetitive and structurally diverse due at least in part to the presence of transposable elements that comprise >85% of its length (Fu & Dooner, 2002; P. S. Schnable et al., 2009). Prior to the array-CGH analyses described in this dissertation our knowledge about the Structural Variation(SV) of non-repetitive regions and the gene space of maize was quite limited. Our results demonstrate that even in non-repetitive regions, maize exhibits high levels of SV, including hundreds of Copy Number Variations (CNVs) and thousands of Presence-Absence variation (PAVs). Similar results were subsequently reported by other groups surveying additional inbred lines (Beló et al., 2010; Swanson-Wagner et al., 2010) or using different technologies such as Whole Genome Shotgun (WGS) re-sequencing (Chia et al., 2012) and RNA-seq (Hansey et al., 2012). In addition to CNVs and PAVs, our analyses and those of others identified Highly Conserved Regions (HCRs). CNVs, PAVs and HCRs are not evenly distributed in the genome.

We identified both SV (such as the ~2.6 Mb loss from chromosome 6 of Mo17 relative to B73) and conserved regions (such as a >10Mb HCR on chromosome 8 and the regions surrounding the *tb1* and *y1* genes). This mosaic structure may be the result of the combined effects of maize's mating system and breeding practices. Open pollination can quickly spread SV throughout a population while breeding practice either artificially select certain loci (and nearby regions), such that in elite germplam certain genomic regions are actually inherited from a single ancestor (identity by descent, IBD) or a very limited number of ancestors.

Many PAVs that are present in one haplotype but absent from another contain intact, expressed, single-copy genes. Due to technological limitations, our CGH analyses could only discover those genes that exist in the reference genome but absent in other inbred lines. To overcome this limitation, Whole Genome Shotgun(WGS) sequencing, RNA-seq and Seq-capture were used to demonstrate that most maize inbred lines contain hundreds of "intact, expressed, single-copy genes" that are not present in the reference genome (Lai et al., 2010; Rustenholz, Ying and Schnable, unpublished data). Over 70% of the CNVs and PAVs were shared among multiple lines. The majority of these shared CNVs and PAVs were observed in both maize and teosinte, suggesting that these variants existed before domestication and that there is no strong selection acting against them (Rustenholz, Ying and Schnable, unpublished data). Some of those common PAV genes co-segregate with major heterotic groups (Hansey et al., 2012), which is at least consistent with the hypothesis that they may contribute to heterosis. A large scale association study

showed that SVs are enriched at loci associated with important traits (Chia et al., 2012), suggesting that some of these SVs may actually be the causative effect of the diversity of those traits.

Although array-CGH based CNV/PAV detection achieved great success in the past 10 years as compared to karyotyping methods, the information that array-CGH can reveal is still limited. The probes of microarray are designed based on known sequence. The array-CGH strategy therefore is not suitable for *de novo* discovery of PAVs. Also, the complex relationship between the copy number and hybridization signals makes it impossible to determine absolute copy number. In this respect Next Generation Sequencing (NGS) technologies has advantages for detecting and characterizing SV(Alkan, Coe, & Eichler, 2011; Medvedev, Fiume, Dzamba, Smith, & Brudno, 2010). It can not only detect SNPs and small INDELs, but it also has the potential to detect other genomic aberration such as CNVs and PAVs (Xie & Tammi, 2009) and even more complex genomic rearrangements (Chen et al., 2009). For NGS-based SV discovery, genomic DNA is sheared into fragments, which are then partially sequenced (36-100 bp) from one end or more typically both ends. DNA sequences are mapped back to the reference genome or *de-novo* assembled into contigs. Different types of sequence information such as read depth (RD), "polymorphisms" among paralogs ("paramorphisms")(Emrich et al., 2007), pair end/mate pair length and aligned reads with extra large gap (split reads) are retrieved and combined with computational/statistical models to identify SVs(Rong Shen, 2012). Unlike array-CGH, which can only allow us to approximately define the region of SV, NGS-based methods can determine the exact boundaries of SV ("breakpoints"). With the relatively higher signal-to-noise ratio of NGS-based methods, attempts have been made to estimate absolute copy number of CNV in each sample and reconstruct the structure of each copy rather than only relative gain or loss(Medvedev et al., 2010).

In combination with recent advances in characterizing maize genome diversity, we estimate that the maize B73 reference genome may be missing thousands of genes relative to the entire gene space of *Zea*. Systematic surveys should be performed on non-reference inbreds to identify genes that are missing from the reference genome. The "Zeanome", a near-complete set of genes present in *Zea*, is being defined using existing genomic sequences and newly discovered gene sequence data(P. S. ;Bren. B. N. L. Schnable, 2011).

It was recently demonstrated in humans that as a consequence of a lack of linkage disequilibrium (LD) many CNVs cannot be tagged well by nearby SNPs (Conrad et al., 2010). Highly efficient and cost saving genotyping methods must be developed for high throughput genotyping of these newly found genes in large populations such as the maize Nested Association Mapping (NAM) panel. The array-CGH based genotyping platform we used in chapter 3 is one candidate. But newly developed sequence based genotyping methods such as genotyping by sequencing (GBS) (Elshire et al., 2011), restriction-site-

associated DNA sequencing (RAD-seq)(Baird et al., 2008) and reduced-representation libraries(RRLs) (Tassell et al., 2008)may be better choices (for review see Davey et al., 2011). With accurate, high-throughput genotyping methods, SVs can be genotyped and over the next few years used to test the hypothesis that SV contributes to variation in phenotypic traits.

The underlying mechanisms for the origin of most SVs are not clear yet. In the human genome most CNV formation is mediated by non-allelic homologous recombination (NAHR), or the corresponding expansion or extraction of variable numbers of tandem repeats (VNTR). But transposable elements such as helitrons may play a more important role in the maize genome (Lai, Li, Messing, & Dooner, 2005).

,

References

Alkan, C., Coe, B. P., & Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nature reviews. Genetics*, *12*(5), 363–76. doi:10.1038/nrg2958

Baird, N. a, Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. a, Selker, E. U., et al. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS one*, *3*(10), e3376. doi:10.1371/journal.pone.0003376

Beló, A., Beatty, M. K., Hondred, D., Fengler, K. a, Li, B., & Rafalski, A. (2010). Allelic genome structural variations in maize detected by array comparative genome hybridization. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*, *120*(2), 355–67. doi:10.1007/s00122-009-1128-9

Chen, K., Wallis, J. W., McLellan, M. D., Larson, D. E., Kalicki, J. M., Pohl, C. S., McGrath, S. D., et al. (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature methods*, *6*(9), 677–81. doi:10.1038/nmeth.1363

Chia, J.-M., Song, C., Bradbury, P. J., Costich, D., de Leon, N., Doebley, J., Elshire, R. J., et al. (2012). Maize HapMap2 identifies extant variation from a genome in flux. *Nature genetics*, *44*(7), 803–7. doi:10.1038/ng.2313

Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., et al. (2010). Origins and functional impact of copy number variation in the human genome. *Nature*, *464*(7289), 704–12. doi:10.1038/nature08516

Davey, J. W., Hohenlohe, P. a, Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature reviews. Genetics*, *12*(7), 499–510. doi:10.1038/nrg3012

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. a, Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS one*, *6*(5), e19379. doi:10.1371/journal.pone.0019379

Emrich, S. J., Li, L., Wen, T.-J., Yandeau-Nelson, M. D., Fu, Y., Guo, L., Chou, H.-H., et al. (2007). Nearly identical paralogs: implications for maize (Zea mays L.) genome evolution. *Genetics*, *175*(1), 429–39. doi:10.1534/genetics.106.064006

Fu, H., & Dooner, H. K. (2002). Intraspecific violation of genetic colinearity and its implications in maize. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(14), 9573–8. doi:10.1073/pnas.132259199

Hansey, C. N., Vaillancourt, B., Sekhon, R. S., de Leon, N., Kaeppler, S. M., & Buell, C. R. (2012). Maize (Zea mays L.) genome diversity as revealed by RNA-sequencing. *PloS one*, *7*(3), e33071. doi:10.1371/journal.pone.0033071

Lai, J., Li, R., Xu, X., Jin, W., Xu, M., Zhao, H., Xiang, Z., et al. (2010). Genome-wide patterns of genetic variation among elite maize inbred lines. *Nature genetics*, *42*(11), 1027–30. doi:10.1038/ng.684

Lai, J., Li, Y., Messing, J., & Dooner, H. K. (2005). Gene movement by Helitron transposons contributes to the haplotype variability of maize. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(25), 9068–73. doi:10.1073/pnas.0502923102

Medvedev, P., Fiume, M., Dzamba, M., Smith, T., & Brudno, M. (2010). Detecting copy number variation with mated short reads. *Genome research*, *20*(11), 1613–22. doi:10.1101/gr.106344.110

Rong Shen. (2012). *Graphical Model and Algorithm for Detecting DNA Copy Number Variation*.

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., Liang, C., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science (New York, N.Y.)*, *326*(5956), 1112–5. doi:10.1126/science.1178534

Schnable, P. S. ;Bren. B. N. L. (2011). Functional Structural Diversity among Maize Haplotypes. Retrieved from http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=1027527

Swanson-Wagner, R. a, Eichten, S. R., Kumari, S., Tiffin, P., Stein, J. C., Ware, D., & Springer, N. M. (2010). Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome research*, *20*(12), 1689–99. doi:10.1101/gr.109165.110

Tassell, C. P. V., Smith, T. P. L., Matukumalli, L. K., Taylor, J. F., Schnabel, R. D., Lawley, C. T., Haudenschild, C. D., et al. (2008). SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries, *5*(3), 247–252. doi:10.1038/NMETH.1185

Xie, C., & Tammi, M. T. (2009). CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC bioinformatics*, *10*, 80. doi:10.1186/1471-2105-10-80